

DOCUMENT RESUME

ED 434 987

CE 077 213

AUTHOR Easton, Peter A.
TITLE Sharpening Our Tools: Improving Evaluation in Adult and Nonformal Education. UIE Studies 4, 1997.
INSTITUTION United Nations Educational, Scientific, and Cultural Organization, Hamburg (Germany). Inst. for Education.; German Foundation for International Development (DSE), Bonn (Germany).
ISBN ISBN-92-920-1068-6
PUB DATE 1996-00-00
NOTE 352p.
PUB TYPE Guides - Non-Classroom (055)
EDRS PRICE MF01/PC15 Plus Postage.
DESCRIPTORS Administrator Evaluation; *Adult Education; Cost Effectiveness; Educational Assessment; *Evaluation Methods; Evaluators; Foreign Countries; Formative Evaluation; Independent Study; *Nonformal Education; *Professional Continuing Education; Program Effectiveness; *Program Evaluation; Qualitative Research; *Sampling; Student Participation; Systems Analysis

ABSTRACT

This book is designed to assist evaluators of nonformal education (NFE) and adult education programs in their continuing professional education and in the task of improving their capacities for genuinely formative, informative, and participatory assessment of educational practice. Fourteen chapters are organized into five sections. Section 1 contains an introduction that discusses the topics to be covered. Section 2 includes three chapters that address some of the conceptual and theoretical issues underlying evaluation practice: design choices, participatory approaches, and systems thinking. The four chapters in section 3 deal with technical aspects of evaluation design: ensuring the quality of information, developing a unifying matrix of inquiry, understanding sampling, and developing qualitative evaluation designs. The five chapters of section 4 treat several particular methodologies of evaluation that could stand to receive greater emphasis in NFE: first the methods of qualitative data collection and analysis, then administrative evaluation, learning outcomes assessment, impact evaluation, and benefit-cost analysis. The one chapter in section 5 offers some brief conclusions regarding the professional development of NFE evaluators and its relation to broader popular participation in educational planning and evaluation. (Each chapter contains references. An index is appended. Contains 22 figures/tables.) (YLB)

ED 434 987

UIE Studies 4, 1997

SHARPENING OUR TOOLS

Improving Evaluation
in Adult and Nonformal Education

Peter A. Easton

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to
improve reproduction quality

• Points of view or opinions stated in this
document do not necessarily represent
official OERI position or policy

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL HAS
BEEN GRANTED BY

Camclator

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

1

UNESCO Institute for Education
German Foundation for International Development

BEST COPY AVAILABLE

CE 077213

UIE Studies 4, 1997

***S*harpening our Tools**

**Improving Evaluation
in Adult and Nonformal Education**

Peter A. Easton
The Florida State University

A joint publication of the
UNESCO Institute for Education
and the
German Foundation for International Development

The UNESCO Institute for Education, Hamburg, is a legally independent entity. While the programmes of the Institute are established along the lines laid down by the General Conference of UNESCO, the publications of the Institute are issued under its sole responsibility; UNESCO is not responsible for their contents.

The points of view, selection of facts, and opinions expressed are those of the authors and do not necessarily coincide with official positions of the UNESCO Institute for Education, Hamburg.

The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the UNESCO Secretariat concerning the legal status of any country or territory, or its authorities, or concerning the delimitations of the frontiers of any country or territory.

© UNESCO Institute for Education 1996
Feldbrunnenstraße 58
P.O. Box 13 10 23
20110 Hamburg, Germany

ISBN 92 920 1068 6

Printed by
Robert Seeman • Neumann-Reichard-Str. 27-33
Haus 8, I Stock
22041 Hamburg, Germany • Tel 040/645-0003

TABLE OF CONTENTS

<i>List of tables, boxes and figures</i>	<i>ix</i>
<i>Preface</i>	<i>xiii</i>
<i>Acknowledgments</i>	<i>xv</i>

PART I INTRODUCTION

1.	BUILDING COMPETENCE IN EVALUATION	3
	The evolving practice of evaluation	3
	Beyond the basics	4
	Broadening the focus	6

PART II CONCEPTUAL DIMENSIONS OF EVALUATION DESIGN

2.	EVALUATION DESIGN AND INFORMATION USE	13
	Evaluation design as a learning process	14
	Macro and micro design	19
	Product and process dimensions	22
	Centralized and decentralized designs	23
	Looking forward	26
3.	EVALUATION AND PARTICIPATION	28
	Participatory evaluation: By and for whom?	29
	Participatory evaluation: How?	33
	Developing a choreography	40
	Conclusion	52

4.	SYSTEMS THINKING AND PROGRAM THEORY	56
	Systems modeling of NFE programs	57
	Program theory	67
	Drawing up the balance sheet	75
	Conclusions: Learning to paint the whole picture	80
	 PART III TECHNICAL DIMENSIONS OF EVALUATION DESIGN	
5.	DATA QUALITY	85
	The philosophical stance of evaluators	85
	Consequences for evaluation	87
	Issues of accuracy	89
	The concept of workable truth	100
	Conclusions	103
6.	INDICATORS AND INSIGHT	107
	The overall framework	108
	Strong and weak indicators	111
	Output and process indicators	119
	Sources of information: Relating MIS and evaluation	121
	An information generation matrix	123
	Evaluation "mosaics" and decentralized MIS	126
	Conclusions	127
7.	SAMPLING STRATEGIES	130
	The varied dimensions of sampling	131
	Quantitative sampling	132
	Qualitative sampling strategies	148
	Mixed methods	153

Tables of contents

v

Other important dimensions of sampling	154
Conclusion: The universality of sampling concerns	157

8. **QUALITATIVE EVALUATION
METHODOLOGY: A DESIGN ON MEANING** 159

Who defines the data?	160
Opening up the black box	162
Thick description	163
Meaning patterns	166
The issue of "objectivity"	171
Types of qualitative inquiry	173
Research design consideration	176
Conclusion: The strength of soft designs	178

**PART IV
PARTICULAR METHODOLOGIES
OF EVALUATION**

9. **THE PRACTICE OF QUALITATIVE
EVALUATION: ARTISTRY AND
TECHNIQUE** 183

Collecting and managing data in qualitative designs	183
Data analysis	191
Steps in deciphering meaning	193
Putting it all together	203

10. **ADMINISTRATIVE EVALUATION** 207

Defining administrative output	208
Conducting an administrative evaluation	213
Nurturing a professional bureaucracy	219
Conclusions	221

11.	SELF-DIRECTED LEARNER ASSESSMENT	225
	An example of participatory process	226
	Learner assessment of literacy results	227
	Post-literacy assessment	237
	Conclusions	239
12.	IMPACT EVALUATION	242
	Impact as a realm of program operation	243
	Reconstituting the program strategy	244
	Defining the program's effects	245
	Gauging actual impacts: Rationalistic and naturalistic approaches	251
	Organizing impact evaluation	257
	Conclusion: Participatory evaluation of impact	262
13.	ECONOMIC EVALUATION AND COST-BENEFIT ANALYSIS	266
	The economics of education: Underlying notions	267
	The tools of economic evaluation	277
	Levels of economic analysis	284
	Comparing costs and benefits	286
	Applications to nonformal education	294
	Conclusions: Economic metaphor and educational process	296
	PART V	
	CONCLUSIONS	
14.	THE CONTINUING EDUCATION CHALLENGE IN EVALUATION	303
	Taking the bulls by the horns	303
	Obstacles to enhancing technique and participation	305
	Hope for enhancing technique and participation	307

Tables of contents

vii

Strategies for enhancing technique and participation

310

Envoi: Optimism, patience, and perseverance

314

APPENDICES

INDEX

317

LISTS OF FIGURES, TABLES AND BOXES

FIGURES

2.1	Cycle of activities in educational program implementation	17
2.2	Compound pattern of evaluation design	19
3.1	Graphical representation of "participation space"	32
3.2	Concentric design in participatory evaluation	48
3.3	One scheme for linking participatory planning and participatory evaluation in NFE programs	51
4.1	Adaptation of systems model for evaluation purposes	58
4.2	System model applied to instructional programs	61
4.3	Systems representation of an NFBE program	65
4.4	Flow diagram of evaluation comparisons	77
4.5	Stufflebeam scheme of evaluation comparisons	79
10.1	Schematic representation of domain of administrative evaluation	211
11.1	Self-evaluation booklet from Burkina Faso	231
	11.1a. Cover	231
	11.1b. Reading test in More	233
11.2	Self evaluation booklet from Morocco	233
12.1	Causal scheme for relations between training and improved agricultural yields	252
12.2	Multiple dimensions of post-literacy	264

TABLES

3.1	Framework for design of a participatory strategy	39
3.2	"Map" of stakeholders in school governance decisions	41
4.1	Tabular presentation of the program or intervention strategy of an NFBE program	66
6.1	Information generation matrix: Instructional domain	124
7.1	Sample size required to estimate a proportion whose true value is near 50%.	140
7.2	Sample size required to estimate the average age of a population with an error probability of 0.05 (in 95 out of 100 cases)	140
7.3	Example of stratified sampling frame	144

BOXES

2.1	An example of macro design: Formative evaluation of the IIBE Project in Haiti	20
2.2	An example of micro design: Assessment of mental math skills among local merchants in Niger.	21
3.1	A participatory vignette: Development of self-managed farm cooperatives in Niger	34
3.2	The importance of lateral communication: Nonformal radio education programs in the Republic of Mali	45
3.3	First steps in participatory planning: Two case examples	50

4.1	Programs that "just grew"	68
4.2	Building program theory: The Educational Restructuring (BOETP) Project in Thailand	72
4.3	USAID's logframe matrix	74
5.1	What differences it makes: The example of educational sector assessment	89
5.2	Misunderstanding the nature of a sample: The 1936 Presidential election poll in the United States	93
5.3	Triple sources of error in evaluation data: An example from training in oral rehydration therapy (ORT)	95
5.4	Imaginary statistics in Haiti	102
6.1	Indicator choice in evaluation of Ghanaian literacy	114
7.1	Theoretical sampling: Lessons from Machiavelli's the Prince	156
8.1	What do you count? Or how assumptions shape numbers	160
8.2	Thick description: An example from a Malian literacy evaluation	165
8.3	Elite sampling in Morocco: Picking atypical situations	177
9.1	Ethical dilemmas in evaluation: The "bee be min don" effect in Mali	191
9.2	The payoff to good listening: Local pronunciations of the word "cooperative" in Mali	195
9.3	What's in a school? Definitions of educational investment in Haiti	196
9.4	Example of qualitative coding: Codes from a study of job training programs (Berg 1995: 180)	197

9.5	Example of qualitative fieldnotes	199
9.6	Member checks in participatory research and evaluation: The challenge of "restitution" in Senegal	201
10.1	Administrative self-evaluation: "Total Quality Management" at Johnsonville Foods	214
10.2	Evaluating organizational objectives: "Visioning" in a graduate School of Education	216
11.1	Self-assessment in other arenas: Agricultural projects in the Republic of Mali	235
12.1	Brainstorming program impact: Assessing the effects of "service learning" in the University	250
12.2	Tracer studies - no problem: The census of new literates in Niger	257
13.1	Qualitative benefit-cost evaluation in Mali: "Rentabilite" and "rentabilisation"	276
13.2	Rate of return analysis: In Haiti and Botswana it all depends . . .	293

PREFACE

This book grew out of the author's participation in a series of seminars and workshops on monitoring and evaluation of nonformal basic education, organized by the UNESCO Institute for Education and the German Foundation for International Development as part of a project designed to follow up on the 1990 World Conference on Education for All.

The title, *Sharpening our Tools*, translates well what the book sets out to achieve and admirably accomplishes. Addressed to evaluators of adult and nonformal education programs and others interested in their outcomes who already possess the basic skills of that profession, it presents a range of new tools for evaluation and hones them to a fine edge. The book discusses the uses of those techniques, the various evaluation designs in which they are deployed, and alternate ways of interpreting their results. The reader is also encouraged to think more broadly and deeply about the evaluation process itself, its aims and its theoretical underpinnings.

The capacity to evaluate education programs and draw lessons from their successes and failures is an essential component of educational development, and Prof. Easton's book is an important contribution to reinforcing such capacity in the vital area of adult and nonformal education. The UNESCO Institute for Education could not have made it available to those interested in nonformal, adult and basic education around the world without the generous support of the German Foundation for International Development, whom we wish to thank most warmly.

Christopher McIntosh
UNESCO Institute for Education
Hamburg

ACKNOWLEDGMENTS

One incurs many debts while writing a book such as this, particularly given the need to tend, as we all do, to multiple other occupations.

The first series of debts and expressions of gratitude is owed to the sponsors and organizers of the seminars for evaluators of nonformal basic education produced jointly by the UNESCO Institute for Education (UIE) and the German Foundation for International Development (DSE). It was in the course of these seminars and workshops that I first prepared many of the components of the book, and the reactions of participants and fellow facilitators to my presentations were extremely helpful in sharpening my ideas. Particular thanks are due to Dr. Adama Ouane, formerly of UIE and now with UNESCO in Paris, and to Dr. Josef Müller (since retired) and Ms Anja Frings of DSE, who not only accepted my request to publish a book developed from the presentations made to the seminars but then demonstrated extreme forbearance and goodwill during its long gestation.

I wish also to express my appreciation for the different kinds of support received from colleagues and students at Florida State University, who suffered with the lapses in my contributions to academic affairs that followed from preoccupation with the book and who offered helpful criticism and input at a variety of junctures. Dr. Shira Birnbaum, in particular, took it upon herself to read the entire text with care and to make a wealth of valuable suggestions on content and form. Dr. George Papagiannis examined the chapters on qualitative evaluation and offered very helpful critique. On the production side of the coin, Ms. Chari Kingsbury, Mr. Charles Kingsbury and Ms. Chandell Ingle gave major assistance with layout and typesetting, while Mr. Richard Gibson of Open Options International provided invaluable graphics services.

Beyond these immediate colleagues lie the many adult and nonformal educators with whom I have worked over the years, particularly in West Africa, some no longer with us, many still performing minor miracles with little in the way of resources. The frequent use of African proverbs to illustrate and deepen points in the technical discussions to follow gives testimony to what I owe in both

personal and professional terms to my close association with them and to my years in the friendly confines of the Sahel.

As always, though, the greatest debts are those owed to family. My wife Zohre, our children Naseemeh and Naveed, their grandmother Mohtaram Khami and my own mother, Esther Ward Easton, both bolstered my courage to finish the task I had commenced and accepted numerous paternal lapses and absences until it was completed. They are truly co-authors of the book and to them it is dedicated with much love and appreciation.

Part I
Introduction

CHAPTER 1

BUILDING COMPETENCE IN EVALUATION

This book is addressed to evaluators of nonformal education (NFE) and of adult education programs, and to those interested in or responsible for their work. It is designed to assist these staff in their own continuing professional education and in the task of continually improving their capacities for genuinely formative, informative and participatory assessment of educational practice.

A number of good basic texts on educational evaluation in developing countries is now fairly widely available¹. Understandably, however, such books do not and cannot go much beyond presenting the fundamentals of evaluation methodology: the different types of evaluation, basic design options, the organization of evaluation work, and a brief glimpse of technical foundations. This text is intended to take the practicing evaluator the next logical step - that is, to broach somewhat more advanced topics that are of critical importance to successful evaluation and to assist evaluators in obtaining the resources and support they need in order to upgrade their practice.

The evolving practice of evaluation

From its origins as a means of accountability required by foreign and national-level donors, educational evaluation has become an institutionalized practice within national education Ministries, their regional branches and, to an increasing extent, the headquarters and field offices of the great variety of nongovernmental organizations and community-based associations which are responsible for much of the actual delivery of nonformal education services. It has also evolved from a primarily "summative" exercise - that is, one devoted essentially to making end-of-period judgments about the success or failure of programs - to a much more "formative" one, concerned both with providing regular feedback to local and national decision makers about the course of program implementation and with generating insight about how to improve program effectiveness.

Along with this change in orientation has fortunately gone some reinforcement of the numbers and capacities of personnel specialized in evaluation. To date, however, relatively little advanced training in evaluation is available in universities throughout the developing world, or even, for that matter, in the industrialized one. Evaluators must make use of whatever research, instructional and/or social scientific training they may have received, plus the experience acquired by facing the challenges of program evaluation in the field, in order to develop the competencies required by their new "profession." Fortunately a few international institutes like the UNESCO Institute for Education (UIE) and the German Foundation for International Development (DSE) have launched a concerted effort to support and upgrade the capacities of nonformal education evaluators in the field; but that undertaking is small in scale, relatively speaking, and its promoters are still seeking ways for optimally dovetailing such outside resources with the particular internal momentum of NFE practice in each interested country.

This book is designed to fit into that emerging effort. Essentially self-instructional written material like that presented here obviously has its limitations and is only one element of what needs to be a multi-pronged strategy of professional and technical development for evaluation personnel. The other components of this strategy potentially include cycles of university training, field-based training sessions, organized exchanges among countries with similar evaluation concerns and greater mutual support among evaluators in different sectors within each country. We shall return to the question of strategies for supporting the continuing education and professional development of NFE evaluators in the concluding chapter of this book. For the moment, let's simply note that one critical step in accomplishing this task doubtlessly resides in promotion of a better appreciation among decision-makers themselves of the importance of good evaluation practice, of the substantial technical content of the field, and of the legitimate training needs of their evaluation staff. Hopefully, this book will help to serve that purpose as well.

Beyond the basics

In writing a book such as this, one must decide *which* aspects of evaluation methodology to treat in the relatively limited space of a single text. There are so many potential dimensions that one might want to develop

that selection criteria must be applied. Some guidance is provided by the nature of the audience to which the chapters that follow are principally addressed: namely, practicing evaluators (and interested associates or superiors) who already have - either by previous training, self-directed learning, field experience or some combination of these - a good basic understanding of evaluation practice and methodology, but who desire to deepen their skills and understanding in order to make more of the great opportunities for program improvement that evaluation offers. Given this orientation and the availability of a number of very good basic evaluation texts, we shall focus instead on what one might call "intermediary" material and assume that basic ideas of evaluation planning, data collection and descriptive analysis are already part of the reader's professional baggage.

Three other criteria have guided the selection of topics dealt with in the pages to follow. First come the felt needs of evaluation personnel themselves. Thanks to the years of experience that the UNESCO Institute for Education (UIE) and the German Foundation for International Development (DSE) have accumulated in short-term training of evaluation personnel from Africa and Asia, their staff have had a number of opportunities to assess the present capacities of NFE evaluation personnel and to discuss with them the kinds of practical training they most need or desire. The choice of subjects in this book reflects to a considerable extent the topics most frequently requested by participants in those sessions.

The second selection criterion might appear to contradict the practical focus of the first: given page limitations, there are good reasons for devoting nearly as much space to the *theoretical underpinnings* of evaluation practice as to specific techniques and methodologies. The advantage of good theory is both that it is applicable to multiple situations and that it helps the evaluator to focus on underlying processes as well as on their surface manifestations. Among the theoretical perspectives developed in the following pages are those involved in the systems analysis of NFE programs, in identifying the determinants of data quality, in probing the nature of cause-effect relations in social change efforts and in using the concept of economic efficiency for educational assessment.

The third criterion adopted here might also seem paradoxical, yet is one of the most important: whatever is done for the training and professional development of NFE evaluators should, this author believes, lend

itself to ease of retransmission to field staff and to program beneficiaries themselves. The job of evaluators, and particularly of those concerned with adult education and NFE programs, which typically place a strong emphasis on beneficiary participation and empowerment, is as much to teach and enable *others* to evaluate their own practice as to evaluate it themselves. As Bhola points out (1991: 65), one key goal of good evaluation practice is to help nurture within existing cultures - traditional and administrative - a "culture of information" in which accurate perceptions of what is happening in a program and of the effects it is having are freely generated, freely shared and used as the currency of collective decision-making. For such an environmental transformation to take place, evaluators must not just acquire new skills but must also know how to discern their basic theoretical underpinnings and how to "empower" other program stakeholders with this knowledge. As always in education, the real criterion of mastery is the ability to penetrate to the essentially simple and humanly relevant heart of technical matters and to share these core notions with others. Evaluation competence is therefore by definition a participatory affair, and a concern for promoting such dissemination of knowledge and skill will enter into the choice of topics in this book.

Broadening the focus

What kind of menu of topics do these criteria suggest? Since whatever mix is adopted is sure to be in a sense little more than a "sprinkling," I will try to spread topics across the span of program operations in order to constitute a better sampling of some of the directions that further training or self-instruction in evaluation may take. Given the degree to which instructional practice has been (understandably) the prime focus of evaluation practice in NFE, I will tend moreover to emphasize better analysis and assessment of all the factors that lie "upstream and downstream" from the courses themselves and that so often determine their effectiveness. Feelings of inadequacy in regard to this necessarily thin choice of topics can be somewhat assuaged if we remember the underlying themes of the book:

- that continued professional development and training of NFBE evaluators is both critically important and eminently possible;

- that such training must pay as much attention to deepened theoretical understanding of program processes as it does to flashy techniques; and
- that the bottom-line criterion is how well this improved practice can be explained to, and spread among, the many other stakeholders of NFBE programs.

Hopefully some progress will be made in addressing these themes in the chapters to follow, even if the coverage of "intermediate" aspects of evaluation methodology is necessarily a little idiosyncratic and spotty!

The chapters of the book are organized into five sections. The first contains this introduction. Section Two includes three chapters that address some of the *conceptual and theoretical issues* underlying evaluation practice: design choices, participatory approaches, and systems thinking. Section Three deals with *technical aspects of evaluation design*: ensuring the quality of information, developing a unifying matrix of inquiry, understanding sampling, and developing qualitative evaluation designs. Section Four treats several *particular methodologies of evaluation* that could stand to receive greater emphasis in NFE: first the methods of qualitative data collection and analysis, then administrative evaluation, learning outcomes assessment, impact evaluation, and benefit-cost analysis. The final section of the book offers some brief conclusions regarding the *professional development of NFE evaluators* and its relation to broader popular participation in educational planning and evaluation.

Notes

¹ See in particular those listed as references at the end of this chapter.

References

- African Development Foundation. 1995. *Participatory Evaluation in Practice*. Washington, DC: The Foundation.
- Ahmed, M. 1975. *The Economics of Nonformal Education: Resources, Costs, and Benefits*. New York: Praeger.

- Berk, R.A. 1990. *Thinking About Program Evaluation*. Newbury Park, CA: Sage Publications.
- Bhola, H. 1990. *Evaluation "Literacy for Development" Projects, Programs, and Campaigns: Evaluation Planning, Design and Implementation, and Utilization of Evaluation Results*. Germany: UNESCO Institute for Education.
- Bhola, H.S. 1979. *Evaluating Functional Literacy*. Amersham: Hulton for the International Institute for Adult Literacy Methods.
- Bingham, R.D. 1989. *Evaluation in Practice: A Methodological Approach*. New York: Longman.
- Casley, D.J. 1988. *The Collection, Analysis, and Use of Monitoring and Evaluation Data*. Baltimore: The John Hopkins University Press.
- Choudary, A. and R. Tandon. 1992. *Participatory Evaluation: Issues and Concerns*. New Delhi, India: Society for Participatory Research in Asia.
- Easton, P.A. 1984. *L'éducation des adultes en Afrique noire: Manuel d'auto-évaluation assistée*. (Adult Education in Black Africa: Self-Assessment Manual). 2 vols. Paris: Editions Karthala.
- Fetterman, D.M. 1996. *Empowerment Evaluation: Knowledge and Tools for Self-Assessment and Accountability*. Thousand Oaks, CA: Sage Publications.
- King, J.A. 1987. *How to Assess Program Implementation*. Newbury Park: CA: Sage Publications.
- Levin, H.M. 1983. *Cost-Effectiveness: A Primer*. Beverly Hills, CA: Sage Publications.
- OECD. 1988. *Evaluation in Developing Countries: A Step in a Dialogue*. Paris: OECD Publications and Information Centre.
- Patton, M.Q. 1990. *Qualitative Evaluation and Research Methods*. Newbury Park, CA: Sage Publications.
- Payne, D.A. 1994. *Designing Educational Project and Program Evaluations: a Practical Overview Based on Research and Experience*. Boston: Kluwer Academic Publishers.
- Shadish, W.R. Jr., and Thomas D. Cook and Laura C. Leviton. 1991. *Foundation of Program Evaluation: Theories of Practice*. Newbury Park: Sage Publications.

Tilburg, P.V. and J. de Haan (eds.) 1995. *Controlling Development: Systems of Monitoring & Evaluation and Management Information for Project Planning in Development Countries*. Tilburg: Tilburg University Press.

Valdez, J.J. 1994. *Monitoring and Evaluating Social Programs in Developing Countries: A Handbook for Policymakers, Managers, and Researchers*. Washington, DC: World Bank.

Part II

Conceptual Bases of Evaluation Design

CHAPTER 2

EVALUATION DESIGN AND INFORMATION USE

Much of the skill and experience needed to improve evaluations and to increase their utility for nonformal education programs can be boiled down to a question of "design." In the chapter to follow, we will examine some of the dimensions of evaluation design that practitioners learn progressively to master. These matters constitute the first focus of the book precisely because, it is often maintained, such design issues are to an evaluation more or less what a road map is to a trip or an architectural plan and a good foundation are to a house: they guide the work that is to follow and play a major role in determining its worth.

The analogy can be taken as true on one major condition: that the word "design" be understood to refer to a "process" as much as to a "product." David Korten has frequently made this point with respect to project plans (e.g., Korten 1980). He contrasts a "blueprint" approach to planning with a "learning process" approach. The blueprint analogy - essentially a notion borrowed from engineering - implies developing initial specifications about virtually everything that will be done and then requiring strict adherence to them in all phases of implementation. This is usually a recipe for failure in human service undertakings like education where cause and effect are not always clear and responsiveness to human variables is essential. What Korten calls the "learning process" approach, on the other hand, is based on the recognition that in social service projects appropriate objectives and methods are never entirely clear and ordained at the outset, since they are often situation-specific and may be viewed differently by the different parties concerned with the project.

Good project development is therefore much more of *an iterative and interactive process* than the engineering analogy suggests. One develops an initial plan, then modifies and refines it on the basis both of reactions from stakeholders and of initial results in the field. In a sense, the final design is as much a *result* of the project as it is a necessary precondition for it. In this approach it is consequently as important carefully to provide for the ways in which project design can be debated among stakeholders and modified in response to what is learned from the

field as it is to develop and polish all specifications at the outset. Such a conception of program and project development *puts a very high premium on evaluation*. To provide the assessments of interim results and the feedback from stakeholders that will serve as the stimulus and the necessary input for modifying and improving initial plans, decent evaluation is obviously a must. In short, a "learning process" approach is inconceivable without a well-developed evaluation capacity.

Evaluation design as a learning process

If evaluation in its various forms is one of the primary means by which organizations shift from a "blueprint" approach regarding program development to more of a "learning process" one, can the same perspective be applied to the development of evaluation designs themselves? Must evaluations be structured like blueprints or can they take form and change form in the process of implementation? What degree of flexibility should evaluators allow and encourage in this regard? The answer, I think, is "a great deal," though paradoxically it requires much more mastery of technique to be flexible than it does to be rigid. This is because flexibility does not mean formlessness or absence of content, but rather *options*. A "flexible" technician is one who recognizes that there are several alternative ways for accomplishing given ends, who knows how to identify them, and who has a sufficient mastery of the entire spectrum of approaches and of their underlying principles to explain these alternatives to stakeholder groups and to assist *them* in understanding the implications of the different options and in making the choice.¹

Donald Schon has called this kind of professional a "reflective practitioner" - one who knows that in human affairs there are no universally applicable formulas and therefore does not invoke his or her expertise or diplomas in order to dictate solutions (Schon 1983). On the contrary, the reflective practitioner is able to approach each situation as unique in many respects, express his or her puzzlement before (and so appreciation of) its unusual and fascinating dimensions, and thereby involve his or her "clients" in actively learning how to resolve the problem. Reflective practitioners accomplish this task by modeling learning behavior - i.e., by showing their clients that the most important skill to exercise in such puzzling situations is the ability to learn in and

from them - and by revealing the reservoir of precedent, accumulated knowledge and usable "templates" (Schon's word for methods that have worked before in similar circumstances) on which professional and lay people alike may draw. In this kind of professional framework, techniques are not prescriptions or formulas that dictate the single best way to design or implement a program. Rather they are templates and prototypes: that is, tools that may help to organize reflection and action, malleable procedures that may be bent and shaped to many different uses.

Such a perspective might seem at first glance to reduce radically the role and importance of evaluation design; but looks are deceiving. The learning process approach in fact *increases* the importance of evaluation design, but changes its meaning. Design is no longer a precise prescription of what is to happen but rather (a) a source of relevant options regarding evaluation development and (b) a process for choosing among them, learning from them and building evaluation plans. At the same time, in a learning process approach it becomes evident that there are important design choices at several levels. They concern not only options regarding the characteristics of individual evaluation studies, but also alternative configurations for the entire evaluation "program." By "evaluation program," I mean the whole set of evaluation and research studies that a project or organization undertakes over time in the interests of improving its operations, and the distribution of roles and responsibilities for accomplishing these evaluative purposes among various units and echelons of the organization.

Developing capacities for collective reflection

In short, evaluation is one of the prime mechanisms, if not the most important one, by which organizations and groups of people *learn*. Individuals within an organization or group may learn a great deal from their experience; but, as Argyris and Schon pointed out some years ago, those lessons will have little impact on collective behavior or upon the effectiveness of programs unless they become part of the "organizational culture" and are incorporated into the models and program theories that govern how resources are used and field activities carried out (Argyris and Schon 1978). Evaluation provides a means of formulating collective lessons of experience and incorporating them into organizational theories of action. Peter Senge has recently developed still further this notion of

evaluation as an organizational learning tool (Senge 1990). He maintains that the only agencies, businesses and associations that will survive the major economic and demographic changes of the turn of the century, at least in industrialized nations, are those "learning organizations" in which employees are regularly engaged in holding up to scrutiny the full range of organizational norms, managerial procedures and production techniques used to accomplish the organization's purposes and are systematically involved in acquiring the new knowledge they need to reshape these procedures on an ongoing basis. And he identifies evaluation as a key instrument in the process.

From this point of view, all species of evaluation are really "formative" in purpose - that is, devoted to improving practice - and only the time frame and administrative setting change as one moves from monitoring to formative assessment and on to evaluation research or summative evaluation. All involve developing the "learning capacities" of the organizations and social groups for whom evaluation is performed and refining their ability to understand their own experience, to learn from it and to improve their practice as a consequence. Evaluation thus fits naturally into the cycle of activities necessary to the implementation of an educational program, portrayed in Figure 2.1. The cycle begins with initial planning and moves through program development to actual field activities, accompanied by appropriate types of monitoring. Data from these activities furnish a basis for evaluation, which provides the information and insight necessary to modify plans and develop new program elements (or revise existing ones) - and the cycle recommences. Research, research and development (the design and testing of new strategies and materials), and collection of management information (MIS) go on continually and feed into all the cyclical processes portrayed.

Of course, in reality the activities in question are not as rigidly sequential as portrayed. In a sense, all are going on simultaneously, and it is more a case of each one coming "front-stage," so to speak, at a given point in the cycle, than it is of the activity being limited to that time single slot. Yet the diagram does summarize an inter-relatedness among these functions that evaluators should keep in mind. Moreover, it becomes evident that evaluation is potentially handmaiden to many other critical activities within the operations of an educational program. The information and insights that it provides fuel planning and program development, guide implementation, and suggest important new

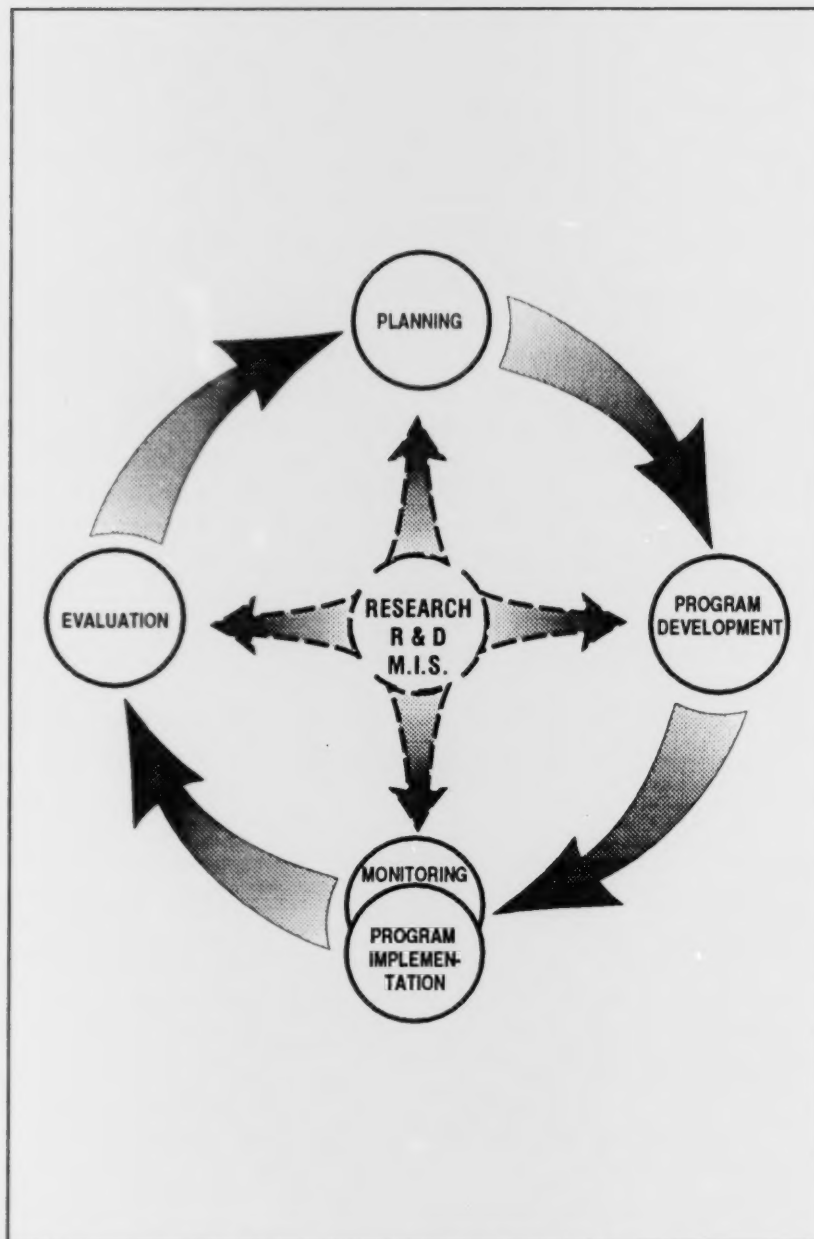


Fig. 2.1 Cycle of activities in educational program implementation

directions for research and for R&D. Together these functions form a whole, and none makes much sense without the others.

Adapting evaluation to its uses

It follows from this line of argument that all evaluation is necessarily heavily "utilization-focused," to borrow the term that Michael Patton popularized a few years back (Patton 1986). Its fundamental purpose is to furnish information that can and will be used by the different parties responsible for executing an educational program in the decisions that they must repeatedly make regarding its design or redesign, implementation, replication and funding. Different kinds of information are needed by the various actors involved, however, at different points in this process. Information to be used for the immediate purposes of monitoring the key "vital signs" of program implementation on a day-by-day or even month-by-month basis is not the same thing as that needed for yearly decisions regarding renewal of the program, and neither of these uses of information is quite the same as the kind involved in deciding what parts of a program to replicate on a larger geographical scale or to apply to totally new projects.

These multiple functions explain why the word "evaluation" as used here does not refer to a one-shot endeavor undertaken to satisfy a single funding or assessment decision, but rather to a whole *program of activities* conducted on an ongoing basis by an educational or social agency and designed for at least three purposes:

- to furnish stakeholders with information of the type they need and at the moment they need it - in order to better understand the workings of the educational undertaking for which they are responsible and to make more productive decisions concerning its administration and renewal;
- to assist them in learning about evaluation and assessment themselves; and
- to build up the reflective and learning capacities of the entire organization to which they belong while increasing its store of relevant insight about its own operations and the opportunities and challenges in its environment.

		DESIGN ASPECT	
		PRODUCT (what)	PROCESS (who, how, when)
DESIGN LEVEL	MACRO (overall)	Yearly plan of evaluation activities	Strategy for developing overall evaluation scheme
	MICRO (single evaluations)	Design documents for individual evaluation studies	Method for generating specifications of individual studies

Fig. 2.2. Compound pattern of evaluation design

Macro and micro designs

Evaluation "design" could therefore be thought of as a compound entity with four categories, portrayed in the two-by-two table of Figure 2.2. On one axis is the distinction between "micro designs," or the plans governing the execution of individual evaluations and evaluation research studies, and "macro designs," or specification of the nature and interrelations of all the varied kinds of evaluation that can go into making up the overall evaluation program of any organization. On the other axis is the distinction between the "product" and the "process" dimensions of evaluation design, or between written prescriptions regarding methodology and focus on the one hand and, on the other, procedures governing how these plans are formulated, shared with all concerned parties and progressively refined and finalized. Let's examine these distinctions a little more closely.

The macro design question concerns decisions about which kinds of evaluation activities to undertake, the relative priority to give to each and the roles that different stakeholder groups will play in each or the way

that these tasks will be distributed among levels of an agency or the different organizations in a network. (An example of macro design appears in Box 2.1.) In short, what is at stake in macro design decisions is the shape or configuration of an overall evaluation program. The elements in this overall design may include the following kinds of studies or data collection procedures:

- *Monitoring and Management Information System (MIS)* activities devoted to assembling on a regular basis key indicators of program or project operations for use in monitoring implementation.
- *Formative evaluation*, which generally involves putting together monitoring data on a periodic basis (e.g., once a year) and supplementing them with additional information on program outcomes and processes. The purpose is to draw conclusions about the interim results of the program and the adequacy of its methodology, and on this basis to suggest corrective modifications in the strategy that are likely to improve its effectiveness.

*Box 2.1. An example of macro design:
Formative evaluation of the IIBE Project in Haiti*

In the late 1980's, the Haitian Foundation of Private Education (FONHEP) developed a comprehensive program of evaluation to provide guidance for its implementation of the Incentives for Improving Basic Education (IIBE) project, an initiative supported by foreign aid and designed to help resource-poor private primary schools in the country's rural and periurban areas improve their instructional quality and administrative efficiency. That evaluation program included a range of different types of evaluation inquiry, each with its own micro design, carried out over different but overlapping periods of the year. Some of its most important components were a system for monitoring program implementation and gathering management information systems (MIS) data, a series of intensive qualitative studies of classroom interactions and home-school relations in selected localities, an ongoing analysis of school finances as part and parcel of the "performance contract" concluded between each school and the project, and a battery of tests carried out twice a year in a sample of schools to track student progress. Results from the different components of the "macro design" were compared and presented yearly to the assembly of the Foundation and to participating schools.

- *Evaluation research*, which typically involves investigating particular critical issues or "knowledge gaps" that have become apparent during the implementation of the project or its interim evaluation, and doing so in a more intensive and research-based mode than is the case for monitoring or formative evaluation.
- *Research and development (R&D)* assessment, which is usually an integral part of any R&D effort undertaken to develop new methods or materials for use in an educational program. In R&D, there is often a very careful alternation between development of prototype methods or materials, their trial application in the field, evaluation of these trials, and progressive refinement of the innovation.
- *Summative evaluation* taking place at the end of the project or program, or of a given multi-year cycle of funding, and designed to draw whatever lessons can be gleaned from the overall results of the activity, to estimate the longer-term impact that it has had or is likely to have, and to decide about its possible replication.

Box 2.2. An example of micro design:

Assessment of mental math skills among local merchants in Niger

In an attempt to improve the mathematics curriculum of its adult education and literacy offerings and to relate them better to existing needs and competencies in target areas, the Literacy and Adult Education Service (DAEP) of the Republic of Niger carried out in the early 1980s an evaluation research study designed to determine how and how well unschooled adults - particularly traditional merchants accustomed to commercial transactions - managed to calculate sums in their heads without written support. A sample of rural and periurban sites known for their commercial activity was selected. The objectives of the exercise were carefully defined and indicators to of "mental math" performance were detailed.

Next, evaluators designed, pre-tested and then finalized a research procedure and a data collection instrument. The procedure included instructions for recruiting subjects to participate in the study, giving them a graduated series of mental math problems, observing their approach to the task and their performance and then interviewing them about the experience and about their methods for resolving everyday math problems was devised and pre-tested. The resulting study was carried out in the field over a two-month period by central evaluators with the help of regional literacy agents

Within each of these categories, of course, there may be different kinds of data collection strategies: quantitative approaches, qualitative designs, mixed-methods approaches, etc. How they all fit together, alternate in time, and complement and inform each other is part of the artistry of "macro" evaluation design.

The *micro* design question concerns the planning of individual evaluation studies and is the one traditionally associated with the classic issues of research design: what questions to ask, what information to collect, where to collect it, how and when; and how to analyze it, promote use and dissemination of the results, and store the resulting data for ease of later access and use. (An example of micro design appears in Box 2.2.)

Since evaluation is, however, a different activity than research, the design will also be somewhat different. Where research studies often place emphasis on the generalizability of findings (quantitative designs) or the "thickness" and evocative value of descriptions and meaning attributions (qualitative designs), in *evaluation* priority is generally given to the local or program specific pertinence and usefulness of the information gathered. Many of the techniques of research are used in educational evaluation, but they are most often used to somewhat different purposes.

This, too, is a point to which we shall return.

Product and process dimensions

In each of these domains, there is also a "product" side and a "process" side to evaluation design. In the macro setting, the product is some sort of (e.g.) yearly or multi-year plan for the evaluation program of an organization, including specification of the priority questions to be answered, the way that information gathering strategies are distributed across different kinds of studies and data collection activities at different points in time, how the results of these different efforts will build a better and broader understanding of critical issues, and how the information produced will be analyzed, disseminated and used by relevant stakeholder groups.

Micro design products, on the other hand, are the design statements for individual studies, like the proposal for a given effort of formative evaluation, the MIS of a particular regional office, or the design of a

single evaluation research activity devoted to elucidating a specific set of problems. The classical format for this sort of single-study design proposal (which it is by no means imperative to follow) includes a statement of the problem or knowledge gap to be addressed, a review of what is already known about the topic, a definition of the specific questions one hopes to answer, and presentation of the methodology proposed for answering them: that is, operationalization of variables or themes of interest, definition of indicators, construction of instruments, establishment of sample, data collection procedures, and data analysis procedures. The design document usually concludes with a careful examination of the resources and time required to carry out the work. The design is the "game plan" for accomplishing the single study in question.

The "process" side of the equation in both cases concerns how these designs are developed, modified and finalized; who takes part and when; what information or training they will need; and so forth. Particularly in the case of participatory evaluations, of which more will be said presently, the entire evaluation undertaking can usefully be seen as a form of teaching and learning for all involved, and the process dimension of evaluation design as the equivalent of a lesson plan. A major evaluation is, from this perspective, a large collective effort at organizational learning; and *how* it is conducted is every bit as important as *what* is investigated. In particular, the larger the evaluation, the more effectively its design may need to be both "decentralized" and coordinated if many actors are to participate and the activity is intended to be both "formative" and instructive. This point merits some additional commentary.

Centralized and decentralized designs

The notion of a "learning process" approach to design introduces one valuable note of flexibility into the domain of evaluation planning, but there is a still more fundamental issue to be faced that is both a hot topic in the current political economy of educational development - not to say of development and governance in general - and a trend of great significance for evaluation practice: the movement for decentralization, for democratization and for privatization of certain social service functions. The methodology and practice of evaluation principally

emerged between 1950 and 1985 in a context of major development programs funded centrally by government ministries or international donors, and its methodology was initially tailored to furnishing these centralized sponsors with means for monitoring the implementation of experimental or large-scale projects, plus reliable information on effects that could be used for subsequent funding decisions. The methodology has evolved considerably since that time, particularly with the introduction of more participatory approaches such as those highlighted in Chapter Three of this volume, and with the progressive shift of responsibility from outside evaluators to evaluation personnel within the concerned organizations. The image that still underlies many texts on evaluation, however, is that of a large and relatively centralized agency that is charged with administering major social intervention projects and needs to draw lessons from their results.

In fact, the situation in the field has changed significantly in the last fifteen years or so. The emergence and increasing importance of non-governmental organizations (NGOs) as social service providers, the growth of what is now termed "civil society," the expansion of the private sector, and the efforts decentralize or at least "deconcentrate" many government functions have all contributed to creating a situation where social service provision is becoming more of a "mosaic" of local, regional and private efforts supplemented and to some extent regulated by national government agencies than the central and public sector monopoly it once largely was. The role of a central agency or major funding organization with respect to evaluation may therefore involve less conceiving and implementing some vast design for ongoing assessment of all related NFE efforts in a country or region than it will assisting a number - if not a myriad - of different organizations with education programs to perform their own evaluations, coordinating to some degree these efforts, and helping to ensure a certain level of comparability in procedures. "Facilitation" is likely henceforth to be a keyword in evaluators' vocabulary, and an aptitude for training others one of the most important skills required of them.

Letting many flowers bloom

Though dyed-in-the-wool proponents of centralized operations and of the unilateral sovereignty of national agencies may shudder to think of the apparent chaos and disorder of multiple organizations pursuing their

own educational interventions, this situation has in fact certain attractions and advantages for evaluators. It is a byword in research and evaluation that the greater the variability in procedures and results, the more opportunity there is to learn from experience and from the data gathered regarding it. At the outside, if the same approach is used everywhere, the evaluator has no way to judge what works best: all one can determine is whether or not the programs were carried out as planned ("compliance checks") and potentially what outcomes have been produced. It is the *differences* among methodologies employed in different settings (and, in evaluation terms, the analysis of the correlation between these differences and variations in results) that provide the surest insight into the relative effectiveness of alternate approaches. In short, evaluation can be a good deal more productive in a somewhat decentralized context where latitude is given for variations in strategy - on the principle, of short-lived renown in revolutionary China, of "letting a hundred flowers bloom" - than it can in a more uniform environment.

Even so it should be noted that the problem of evaluation design does present itself somewhat differently in these "decentralized" circumstances. First, central evaluation personnel may need to engage in some version of what the French used to call "indicative planning": that is, develop schemes that are potentially reproducible, with appropriate modifications, in varied situations by different sorts of local actors and then assist them in adopting and amending them. Second, evaluators may want to think in terms of "segmented" and perhaps "loosely-coupled" designs: inquiry strategies where different roles are assumed by different organizations and great attention is paid then to the synthesis of what may be quite disparate types of work. Third, central staff may be increasingly called upon to be evaluation advocates with, as well as trainers for, decentralized organizations, in order to convince them of the value of good evaluation practice for their own operations and to help them create the necessary capacity. For this to happen, evaluation must not be perceived as essentially a mechanism of central control nor must evaluators appear to be primarily regulatory agents. The trick is rather to make clear the benefits of evaluation as an organizational learning process for the group itself and to de-dramatize its judgmental aspects, on the general principle that "a well-understood failure is worth at least two successes."

The notion of evaluation planning may therefore change meaning and the design of actual evaluations change shape as we advance into an

era of increasing local and private initiative, but the activity of designing evaluations - that is, carefully considering how to capture the lessons of experience that a program has to offer - seems likely to lose none of its importance. On the contrary, evaluators will surely need increased competence in the conceptual and technical areas visited in the remainder of this book, insofar as they will be increasingly responsible not just for designing and implementing assessments of various types themselves, but for helping others to do so in the many new forms of association now coming to birth. Good "reflective practice" - to return to a theme broached at the beginning of this chapter - requires, if anything, more skill and knowledge than traditional expertise.

Looking forward

Different chapters in the book are most relevant to different parts of this compound picture of evaluation design. Chapter Three on participatory evaluation and Chapter Ten on administrative evaluation, for example, deal directly with the process side of the issue; whereas most of Section Three (Chapters Five through Eight) concerns elements of the design "product." The chapters in the book are not so sequentially constructed, in any case, as to require reading in a strict order. Though there are good reasons for starting with all three chapters of this second section, readers may then focus on whatever issue is of most interest to them and develop, on that basis, their own approach to the rest of the text.

Notes

¹ Of course, the move toward process and flexibility can go overboard, leaving no structure or commonality of purpose. For the most part, what we are talking about here is variations on, and adaptations of, some agreed-upon themes.

References

Argyris, C. and Donald Schon. 1978. *Organizational Learning: A Theory of Action Perspective*. Reading, MA: Addison-Wesley Publishing.

Korten, D. 1980. "Community Organization and Rural Development: A Learning Process Approach." *Public Administration Review*: 480-511.

Patton, M. 1986. *Utilization-Focused Evaluation*. 2nd edition. Beverly Hills, CA: Sage Publications.

Schon, D.A. 1983. *The Reflective Practitioner: How Professionals Think in Action*. New York: Basic Books.

Senge, P. 1990. *The Fifth Discipline: The Art and Practice of the Learning Organization*. New York: Doubleday.

CHAPTER 3

EVALUATION AND PARTICIPATION

Concern for improving the professional and technical competence of staff is often considered to be quite distinct from concern for increasing the participation and "empowerment" of program beneficiaries. The two objectives are sometimes even placed in opposition, particularly if the professional development of staff is thought to confer an expertise or status that places them far "above" the person on the street or gives them access to technical realms that common folk just can't understand.

But one can just as well argue the reverse, and that is the position that will be taken in this book. The chapters to follow are based on the principle that concern for improving technical competence in evaluation is, for several good reasons, inseparable from the notion of beneficiary participation and empowerment. First, the objective of any decent evaluation - and especially of the "formative" type that interests us here - is to provide a basis for improving the effectiveness of NFE programs. And one of the surest avenues to that improvement lies in *disseminating more widely the practice of evaluation itself*: that is, in making those who participate in a program increasingly capable of assessing its strengths and weaknesses and of finding new ways for better accomplishing its goals.

Second, evaluation itself often constitutes a very important "entering wedge" for greater popular participation - by staff, beneficiaries, citizens, employees and/or constituents - in the management and planning of social affairs. One of the best ways to begin getting more people involved in decision-making about programs and institutions is to give them central roles in evaluating the work of those programs and institutions. Such participation in evaluation enables people to get the sort of overall grasp of the operation of a social program that is essential to good management, while at the same time helping them to formulate their own views concerning how those efforts could be more effectively implemented. Moreover, developing avenues for greater popular participation in program direction most often requires a substantial evaluation effort, if only because our programs and public agencies are generally structured in very hierarchical ways that make it difficult to realize

anything other than token participation of beneficiaries. In order to envisage "restructuring" these administrations in ways that will permit greater participation, it is therefore generally essential to undertake first a careful evaluation of the situation.

Thirdly, good evaluation is itself nearly synonymous with good training and with continuing professional education. The experience of participating in evaluation almost unfailingly teaches a great deal to all involved and requires of them that they learn new skills and competencies or improve those that they already have. In short, evaluation is itself a form of "adult education"; and since the function of NFE programs is to provide genuine adult education and empowerment opportunities to the maximum number of people who have been deprived of them, it follows that evaluation is one of the best of such opportunities.

Finally, evaluation, particularly in NFE programs, very often comes down to assessing the degree of effective participation of beneficiaries in program operations and benefits. In short, like most of the programs themselves, evaluation is premised on the notion that the underlying goal of NFE is to empower those involved and to give them the skills and self-confidence necessary to take a more active role in directing their own individual and collective destiny. In fact, evaluators often end up playing the role of gadfly and of advocate for local beneficiaries within such programs - that is, they are often the ones who must say, as diplomatically as possible, that the emperor is not fully clothed, that the promises of participation are not always being realized in the field, and that local staff and beneficiaries have good ideas about how this situation might be rectified. Since assessing and promoting participation is such an integral part of the evaluator's task, it follows that he or she will be sooner or later called upon to practice what they preach - i.e., to make evaluation itself as participatory as they claim the program should be!

For all these reasons, we will consider improvements in evaluator competence and increases in beneficiary empowerment and participation in NFE programs as natural complements. Neither can be accomplished without supporting the continuing professional education of evaluation staff. In this chapter, I wish to set the tone for what is to come by examining briefly just what the notion of "participatory evaluation" means and what sort of technical skills are required to make it a reality. Each of the following chapters will then have a strong "participatory" sub-theme: that is, we will be talking not just about the different methodologies and techniques that evaluators may bring to bear on their work, but how each

can be made an occasion for including broader ranges of people in the evaluation enterprise and promoting their own learning. To use the terms introduced in Chapter Two, the problem is to make evaluation as well into a "reflective profession."

Participatory evaluation: by and for whom?

The word "participation" has become increasingly popular in the last thirty years and is currently used in so many ways and with so many different meanings that it is in danger of losing all content. Concerns about participation are obviously relevant to many more areas than the particular field of evaluation and have acquired additional resonance and currency in recent times from their association with the worldwide movement for greater democracy and accountability in government, industry and nonprofit organizations. So what do we mean by the term and what are its particular applications to the practice of evaluation?

Let's start with the first question: what do we mean by "participation"? A definition with some content and "bite" to it might go as follows:

Participation" denotes initiatives through which people acquire increased and broadened responsibility for decision-making, resource allocation, implementation of activities and assessment of their results in areas of immediate concern to their lives, as well as the competence necessary to play these enhanced roles.

By using a definition like this, we make it clear that we are talking about some degree of actual "empowerment" and not just token participation. Increasing participation means giving those concerned both greater authority and greater capacity to determine how resources are allocated, what objectives are set for social programs, what methods will be used to accomplish those objectives, and how the results will be evaluated - as well as greater opportunity to play an active role in all those phases of the work.

The definition proposed above also makes two other things apparent. First, "participation," like "empowerment," is a question of degree: there can be more or less of it in any given activity. It is not an all-or-

nothing question, though the more those in authority oppose it, the more it begins to look like an either/or issue. Political scientists have devised scales of citizen participation in government affairs, for example, which stretch from passive acceptance of what authority dictates, through regular voting in elections, to active campaigning for candidates and referenda, and on to direct assumption of responsibility for local governance (e.g., Arnstein 1969; Mason 1982).

Bernstein (1982) has suggested that the degree of effective participation in an activity can be mapped along three dimensions, which are arguably quite applicable to the issue of participation in evaluation practice as well. The first concerns what might be called the "breadth" of participation, or the number of facets and phases of the work that are opened up to popular decision-making. In the realm of evaluation, this might mean specifying whether those concerned are going to take part only in data collection, or in evaluation planning, evaluation design, related budgetary decisions, instrument development, data analysis, data interpretation, and dissemination or utilization of results as well.

The second dimension concerns the "depth" of participation, or the number of different strata in the organization or institution that are involved: just middle level staff, or local personnel, beneficiaries, and their families or communities as well? Finally, the last dimension concerns (to stick with the physical metaphors) the "height" of participation, or the level of issues in which participants are given a say. Is their decision-making authority restricted to local policy concerns, or does it extend to regional and national ones as well? The combination of these three dimensions is graphically portrayed in Figure 3.1.

The next thing that becomes apparent as we reflect on the definition of participation suggested above is that participation can be bogus as well as real. (In a sense this is simply a further implication of the matter of "degree" just discussed.) Not all that glitters is gold, as the saying goes: many programs and activities are advertised as highly participatory though actual participation of beneficiaries or staff in decision-making is in fact very circumscribed if not purely token. In NFE evaluations in West Africa, we used as a motto a pithy Hausa proverb that says volumes about the difference between advertised intentions and actual results: *Da an ce de kare, tuwo ya yi yawa a gidan biki, ya ce, "Ma gani a kas!"*: "When the dog was told that there was food for everyone at the wedding feast, he replied simply, 'We'll check that out at ground level!'" There can indeed be all kinds of meat and pastry on the tables at

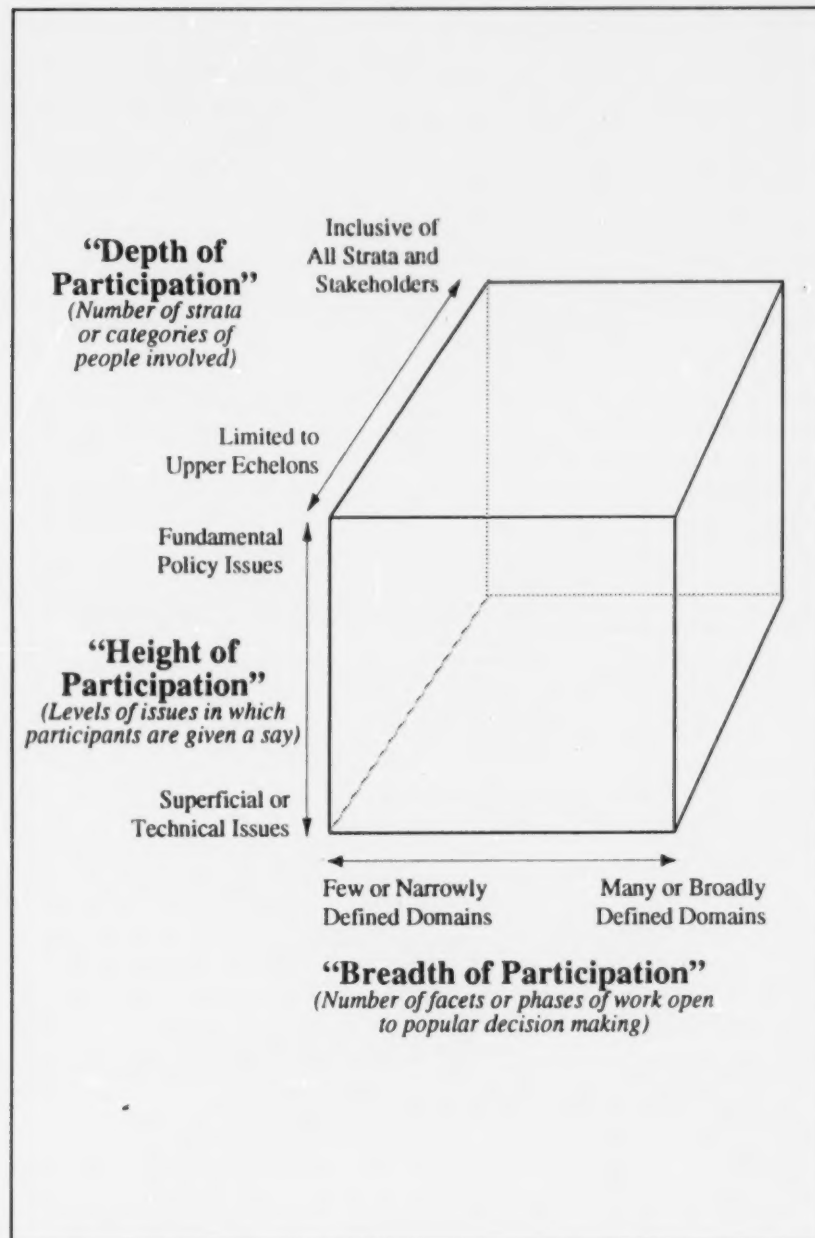


Fig.3.1 Graphical representation of "participation space"

the feast, but if it doesn't get down to the ground, the dog has no part in it. In short, things may appear "nutritious" and highly participatory in publicity or at the headquarters level, but we've got to ask what really transpires in the field.

It is generally no accident that actual participation is so limited. As mentioned above, for a host of historically understandable reasons, our administrative and social structures are for the most part quite hierarchical and "engineering" greater degrees of popular participation in their governance and operations is technically quite an arduous task and politically a sometimes risky one, because it can amount to bringing solidly entrenched interests into the light and into question (Pateman 1970). As Hans Weiler has pointed out, "popular participation" is one of those causes that typically leads to a great deal of divergence between rhetoric and practice, because promising and verbally supporting it has many political benefits, whereas actually carrying it out has high political costs (Weiler 1983: 270). Relatively few people are against some measure of increased popular participation in *theory*. Supporting it in fact may be quite another matter.

The vignette in Box 3.1 brings out this point while supplying an example of potential relevance to evaluation.

Participatory evaluation: How?

If broader participation in evaluation is not incompatible with improved technical performance, how are the two best reconciled? By what methods can evaluation be made truly participatory? What are the techniques involved in promoting participation?

In a sense, this whole book is devoted to answering that question. Each chapter will not only delve into a given technical subdomain of the evaluation "trade" but also suggest how the competencies in question can be widely shared and can serve to stimulate human capacity development throughout the program. But we need to say a few things at the outset about methods for encouraging and developing participation in such technical matters *per se*. They should constitute nothing fundamentally new for adult educators, since the whole field of adult education is an enterprise devoted to increasing people's mastery of their environment and promoting popular participation in technical affairs. But the particular applications to evaluation of this "andragogical" or facilitation task

*Box 3.1. A participatory vignette:
Development of self-managed farm cooperatives in Niger*

In the late 1960's, government rural development agencies in the Republic of Niger undertook an effort to sponsor and support self-managed agricultural cooperatives throughout the country. Participating communities were enabled to assume responsibility for marketing their own agricultural produce and for contracting their own farm credit. In return, they benefited from a proportion of the profit margin on both activities, which they could then invest in other development activities and/or rebate to cooperative members. To assist local people in acquiring the competencies necessary to exercise these new functions, cooperative agency staff and literacy personnel analyzed the different skills needed in market and credit mutual management, organized them by level of difficulty, devised a training program that alternated staged acquisition of these competencies with their actual application, and tested and evaluated it with the assistance of newly literate adults in the target areas. The effort proved both effective and highly motivating for the participants, but ran afoul in a few years of political obstacles because of the shift in power and resources to village communities that it entailed.

Such are often the dilemmas of participatory initiatives. What are the applications of this example to participatory *evaluation*? What are the successive levels of skill that program participants need to acquire in order to play an active role in evaluation? What are the practical applications with which each level of learning can be alternated? And what are the political guarantees to obtain, or precautions to take, so that the initiative can survive?

certainly merit a few initial remarks. These will concern, first, the "learnability" of technical matters; then, the demystification of expertise; and, finally, means for keeping the larger picture in view and for "choreographing" the participation of different categories of stakeholders.

The underlying principle: This ain't so hard to learn

I think that the underlying principle in any effort to promote popular participation in, or oversight of, technical efforts is that the subjects in question are too important to leave to technicians alone and that the central concepts involved are really not that difficult to master, particularly if one is sufficiently motivated. On the contrary, a good analysis of

any technical subject matter generally reveals that it is organized into a hierarchy of levels, all based on the same fundamental principles.¹

Those principles are usually quite simple and understandable, provided that one is able to distinguish them from the (sometimes extraneous) technical virtuosity with which they have become surrounded. What is more, one can generally find in the daily experience of the intended participants certain habits and skills that deal with the same underlying problems. Building on these, it may be possible to make the learning task even simpler. There are even numerous cases where the existing competencies of participants or aspects of "traditional knowledge" in roughly the same domain are at least as refined as the new techniques they are supposed to be mastering, and one can simultaneously improve upon the methodology and adapt it to local applications by incorporating this "local knowledge." A couple of examples will help to make the point.

Consider, for example, the question of sampling methodologies. There are many refinements to sampling methodology, some of which we will consider in Chapter Seven; but the underlying notion is quite simple and intuitively understandable: how to select a relatively small group of people, places or events that is as representative as possible of the whole group (or "population" in statistical language) that one wants to describe but could never hope to survey in the time and with the resources available. Not only is this problem easily understandable; it is in reality one that, for example, field staff and regional supervisors of social programs encounter in their day-to-day experience at least as often as do statisticians.

In my own experience, for example, a regional literacy inspector or NFE supervisor is frequently confronted with the following problem: how to get a relatively good idea of the situation of the literacy or training program in subregion X, given that he or she cannot visit all the centers there in the course of a month? Experienced inspectors know very well that if they visit only centers in urban locations, or those located on main roads, or those near friends' homes, they are probably not going to get a very accurate picture of the overall situation of the program in that subregion. By a good choice of destinations and a good timing of visits, however, it is possible to assess the state of affairs, even if one can not be physically present in more than 10% or 20% of the actual localities in a given time period. And that right there is a sampling problem, plain and simple.

Another example can be drawn from a training session on river blindness for village paramedical personnel that took place in central Mali, West Africa (Belloncle 1983: 261-270). The idea was to initiate these newly literate young people to the diagnosis of the disease and to the identification of the life cycle of the parasite that causes it - some relatively advanced notions that had been the subject of recent research in the medical community. In fact, a medical researcher from the University of Marseilles in France attended the training sessions as a technical resource person. Happily, the organizers had thought to include in the training program a phase of "inventory" of local knowledge related to river blindness, accomplished mainly through interviews and joint sessions with village elders. It turned out that several of those older people knew things about the life cycle of the parasite - from long and close observation, what Levi-Strauss (1966) would call the "science of the concrete" - that the medical researcher from Marseilles had not yet, or had only recently, discovered. Given that solid basis in "local knowledge," the rest of the learning was easily accomplished; and the outside specialists learned a good deal to boot!

The underlying principle, once again, is simply that the fundamentals (and arguably, therefore, the most important part) of technical topics like evaluation methodology, sampling and the etiology of river blindness are both quite understandable and quite parallel to much of our everyday experience. If they appear impenetrable, the problem probably lies in their presentation, in lack of mastery by the technicians responsible, or in someone's conscious or unconscious resolve to keep them under wraps. This perspective implies that two sorts of technique or competence are required in order to promote participation in technical affairs, over and beyond the stubborn faith that such things are possible and that "it ain't so hard to learn."

The first is a capacity to distinguish the underlying principles in any particular technical domain from the progressive levels of refinement that more accomplished practitioners master - or to goad the technicians into doing so!. One can then arrange these competencies into a series of levels in order to define in a cumulative and progressive fashion the skills required to master each and the possible applications of each new level of capacity. This effort amounts to translating the technical domain into a lesson plan - hopefully with the direct collaboration of the technicians themselves - and then saying to them, in effect, "See? It's not so

difficult after all. Let's try helping people learn it step by step and see if that doesn't work!"²

"Defanging" technical expertise

The second "competence" required to promote popular participation in technical affairs is of a more socio-political or diplomatic nature. If it is true that many technical domains (like program evaluation) are not fundamentally as difficult as the public seems to believe, it is also true that this impression is sometimes carefully cultivated by professionals and technicians themselves, or by the institutions that use their services - precisely because it can serve to shield them from public scrutiny or to ensure certain coveted privileges. So the effort to simplify and "demystify" technical matters may run into strong resistance even when it seems both feasible and desirable. One is on contested, if not "mined," terrain. Others may feel that, by trumpeting the "learnability" of technical matters, evaluators or adult educators are poaching on someone else's territory and putting into question the special status of technicians or professionals. There is no simple answer to this dilemma, because higher levels of technical competence *are* very often necessary and useful. It is a question in each case of working out a new *modus operandi* between participants and technicians, one that gives additional competence to the first group and renewed relevance to the second. And great quantities of insight, tact, persistence and social fluency are required to pull it off.

The emphasis placed here on demystifying or "defanging" technical matters and on promoting broader popular access to the tools of the trade should, once again, not be taken as minimizing the importance of technical skill. I do not mean to underestimate either the difficulties involved in increasing the productivity of human effort, or the successive levels of sophistication and training that mastery of any activity, evaluation included, may require. Participatory enterprises need good technical support and critical skill inputs as much as any others. In fact, the fruitful tension between participation and expertise is one of the driving forces of any undertaking (Széll 1989). Both are necessary and the combination of the two is truly dynamic: the challenge is manifestly to increase each without compromising the other. However, perhaps for the socio-political reasons briefly suggested above, the balance in program evaluation and project implementation domains in general seems more likely to be lost through overemphasis on tech-

nique and atrophy of participation than the reverse. Hence the necessity of "the democratic impulse": trying to ensure that broad dissemination of the core skills necessary to understand and monitor the directions of an activity is as well attended to as the more narrowly-based development of its technical refinements. West African wisdom has a humorous and lapidary way of characterizing the minimum knowledge needed to stay conversant with an area. In several of the region's languages, to say one knows at least the rudiments of a language, one says, "At least they can't sell me [i.e. decide my fate] without my knowing it!" In how many areas does the public in our modern world not have that vital minimum?

Table 3.1 illustrates a framework used in a recent study of literacy and decentralization in West Africa (Easton 1995) to analyze the successive levels of competence required to take over management of a technical and socio-economic activity. It also serves to outline the strategy for transferring these responsibilities into local hands, by providing training in alternation with application of new competencies and by indicating the different categories of people or social groups to whom these functions and skills will spread. This sort of scheme should be of interest to evaluators for two reasons: both because it is suggestive of ways to ensure greater participation in evaluation and because the degree of effective "empowerment" of this sort in other domains - and the role that training plays in the process - is often very much a part of what they are called upon to evaluate.

Contextual understanding: putting the program in its place

Another area of critical importance for those wanting to facilitate popular participation in evaluation might be called that of developing and communicating a fully "contextual" understanding of the program to be evaluated - that is, finding a way to situate it with respect to the social habits it is designed to change and the other actors and forces in the environment, and to bring back to critical consciousness the basic postulates on which it is based. We shall dwell on this further in the next chapter when the subject is "systems thinking" and ways to represent the "program theory" or "intervention model" that underlies any NFE activity. But we can at least note here to what degree an interest in promoting participation puts a premium on breadth of vision and critical understanding. In industrial circles, such critical awareness is now

Table 3.1. Framework for design of a participatory strategy

Level of responsibility assumed	Actual technical functions	Knowledge and skills required	Training to be offered	Social groups/categories of persons progressively involved			
				A	B	C	D
I							
II							
III							
IV							
V							

Evaluation and participation

39

sometimes called "double-loop thinking - the ability not just to find ways of doing things better, but to question whether one is doing the right things to begin with and how the very objectives and assumptions of the activity might be reformulated (Argyris and Schon 1978; Scott 1991).

It is also very important to become adept at identifying all the different "stakeholders" (to use the expression now popular in evaluation terminology) who are concerned with the operations of an NFE program, and at understanding the particular "stake" or interest of each. A "stakeholder" is a member of any group whose interests are affected by the program and who therefore wishes to have some say in the way it is handled. The utility of this concept is that it helps to broaden our image of just who is involved in the program, to become more aware of the parties behind-the-scenes or beyond the normal scope of its ministrations who may nonetheless be vitally concerned by what it does, and therefore to take account of their potential interest and include them in the participatory scheme. In a major work on evaluation methods, Guba and Lincoln (1989) in fact maintain that the entire undertaking should be organized around the concerns and perceptions of stakeholder groups. These include not just actors present on the scene in the instructional context (like teachers, learners, local sponsors) but also and manifestly those "upstream" of the activity (like central agency personnel, resource suppliers, foreign donors, etc.) and "downstream" (employers, community leaders, the general public) who are affected by its operations or make use of its results.

One can even think of drawing up a sort of map of stakeholders in order to recognize the different groups that are concerned and the nature of their interest. The one pictured hereafter in Table 3.2, was prepared for a report on educational restructuring in North America and is designed to show all the principal stakeholders in decisions about primary school reform in a typical United States locality.

Developing a choreography: working out the interactions of stakeholders groups

In a participatory evaluation, many more types of people (and, often, many greater numbers of them) are actively involved in inquiry design, data collection, data analysis and the formulation of recommendations than is the case in a traditional external evaluation. The essential objective - as has been noted - is not so much to create some disembodied

Table 3.2. "Map" of stakeholders in school governance decisions [NDR = "not directly represented"]

	FEDERAL	STATE	DISTRICT	LOCAL/SCHOOL
"SUPPLIERS"				
State Executive/legislature	President Congress	Governor Legislature	Mayor City Council	NDR
Education authority	Office of Education	Department of Education	School Board Superintendent	Represented by principal
Other agencies	e.g. Dept. of Health and Human Services	State social service agencies	Community social service agencies	NDR
University/Research Community	National research centers	State University School of education	Teacher education centers	NDR
Interest groups/ Foundations	e.g. Carnegie Foundation	NDR	NDR	NDR
Commercial suppliers	e.g. textbook suppliers	Regional agents	Sales personnel	NDR
Others				
"PRODUCERS"				
Principal	National Association of School Principals	State Association of School Principals	No formal association	Principal's office
Other Adm. Staff	NDR	NDR	NDR	Secretaries, custodial, staff, etc.
Teachers	National teachers' unions	State teachers' unions	District teachers' unions	School faculty
Students	NDR	NDR	NDR	Student body
"CONSUMERS"				
Direct: families	Political representation	Political representation	Political representation	parent-teacher assoc.
Indirect: business	National Association of Manufacturers	State Chamber of Commerce	Local Chamber of Commerce	NDR
Community leaders and members	Political representation	Political representation	Political representation	NDR

objective and external vision of the situation of the program as it is to nurture a new collective consciousness of what is happening and to promote some degree of consensus about how to remedy the problems and realize the potentials of the effort. In a sense, evaluators are just the "midwives" to this new collective consciousness, but in order to create it they need to promote exchanges of information and encounters along new pathways and find or devise the "media" that will make this possible.

In a number of ways, therefore, the design of a participatory evaluation is as much a careful "choreography" of stakeholder groups, information exchanges and events as it is a set of instruments and analysis procedures *per se*. One must decide not just what information to collect, when and where, but also, and very importantly,

- *who* should collect it,
- *who* should analyze it,
- *how and when* the results of analysis by one stakeholder group should be shared with another, and
- *how* information will circulate and feedback be provided so that the points of view of these diverse groups begin, if possible, to converge on some viable conclusions.

That is no mean trick, given the fact that the interests of different stakeholder groups are often at least partly opposed to each other. The evaluator must work with their representatives to find or create common ground. The skills required certainly include, to some degree, those of a good politician or leader - that is, the ability to understand the diverse starting points, values and interests of different groups, to give each a sense of being included and understood, and to create coalitions and bridges among them. One can also speak of three more technical aspects of this task of "choreographing" participatory evaluation and promoting new kinds of information flow: the challenges posed by new communication channels in vertical, lateral and external dimensions. Let's examine each of these briefly.

Vertical communication

The first problem that comes to mind is the one of vertical information flow, since the administrations responsible for NFE programs are often organized in a traditional hierarchical manner. What seems essential here

is creating some kind of "rhythm" by which information can both cascade down and out to the localities where training is taking place *and* flow together and up from these sites to regional or national centers, while being processed and interpreted appropriately at each level. The flow must be two-way and regular (though not necessarily daily or weekly) and it must be cumulative - that is, the communication in each round must take account of what has been transmitted in the previous rounds. In participatory evaluation designs, the flow often starts at the local level (though exactly where the communication originates is less important than the support and initiative given to local beneficiaries for producing their own assessment). Each unit or locality carries out its own data-based self-evaluation, respecting at least in part some agreed-upon standards of form and coverage, and then interprets the results for its own immediate use before transmitting "up the hierarchy" both its conclusions and the supportive data. The subsequent echelon synthesizes the information provided by the local units, verifies or cross-checks as appropriate, adds in data on its own operations, makes a new broader synthesis, draws its own conclusions, and transmits once again.

This same pattern can re-occur from level to level up to whatever constitutes the central echelon in the system, be it a government ministry, a network of NGOs or a company. In a participatory design, it is essential that representatives of "lower" levels participate in the deliberations at the higher ones in order to present their own findings and input and to throw light on interpretations from the point of view of their own more immediate experience of field conditions. In any case, once things are synthesized at the top echelon, the information chain can reverse itself, providing for "restitution" to each level - one after the other or simultaneously - of the results of these broader syntheses and some examples of what people have done and said in other localities. These broader perspectives are usually of great interest to local participants, who are additionally motivated to compare their perceptions with broader ones and with those of comparable units elsewhere precisely because - in a participatory evaluation design - they were in fact the authors of their own evaluation. If evaluation results then serve as the basis for participatory planning, a second cycle of vertical communication commences. (See in this regard Box 3.3 and Figure 3.3 toward the end of the chapter, which illustrate one such vertical flow process.)

Good two-way vertical (or, for that matter, horizontal) communication requires some measure of equality or at least parity and justice

among the parties involved.. Certainly, people of different levels of schooling and social status will be taking part; but in a sense all are equal with respect to the program in question. Those closer to the field probably know the daily reality of its operations better; those closer to the center may have more officially-recognized skills and higher status attributes, and they may also be better able to influence the material and political conditions within which the program as a whole has to operate. Good evaluation requires many ingredients and respect for what each can contribute. Evaluators should remember (and try to get others to recognize) a principle dear to proponents of participatory management: to a significant extent, the "expert" with respect to any job or task or project is the person actually doing it.

Lateral communication

Communication "up and down the hierarchy" is, however, by no means the only type of information exchange in a participatory evaluation. Units and groups of people that share problems or face similar situations need the opportunity to communicate directly with each other and to offer each other mutual support. In fact, participatory evaluation has a tendency to promote a process that seems in any case to be a hallmark of administration in an increasingly democratic age. I am speaking of the relative "flattening" of hierarchies, or the reduction in the number of echelons that separates top management from the field along with shrinking of the vertical distance that separates one level from another and improved lateral flow of information across tiers. The connection probably resides in the fact that the tempering of hierarchies is often associated with policies of decentralization or "management by objectives" (MBO)³, which entail granting a certain degree of budgetary and administrative autonomy to local and regional units; and local evaluation capacity is a key element among the prerequisites for effective decentralization. In short, village and town administrations, local associations or regional offices which have a demonstrated ability to evaluate their own performance according to a set of agreed-upon criteria are much better candidates for decentralization and assumption of increased local responsibility than those not demonstrating such capacity; and since participatory evaluation tends to develop these skills, it is no surprise that it should be associated with the flattening of administrative hierarchies.

In a flatter administrative set-up, there are fewer echelons, and the higher ones serve more to facilitate communication among field units

*Box 3.2: The importance of lateral communication:
Nonformal radio education programs in the Republic of Mali*

Radio education programs were launched in the Republic of Mali in the early 1970s as part and parcel of that country's rural functional literacy efforts and as a supplemental means of feedback and evaluation of the literacy initiatives themselves. Village participants were provided with simple radio and tape recording equipment and encouraged not just to listen to and discuss broadcasts from the program headquarters, but also to send in their own comments, program elements and questions. There was a very strong response from the field, but radio education technicians soon noticed that local participants were even more interested in the opportunity that the program created for communication *among* villages in different parts of the country than they were in the vertical exchange between local and central levels. Most of those submitting tapes were responding to messages and comments of other communities and directly addressing their peers in other parts of the country, and in this way a lively lateral exchange soon developed.

and support their decision-making than they do to give orders and demand obedience to centrally-developed schema. They are the "nodes" in a network of communication and inter-relation, more than the "headquarters" of the army. Under this arrangement, it is significant to note that managerial staff tend to divest themselves of some of their directorial responsibilities and assume "teaching" ones instead. In industries where restructuring and decentralization schemes like "total quality management" have been tried in recent years, it is not at all unusual to hear how the role of middle management has changed from that of supervisor and straw boss to that of "mentor" and "coach."

Whether or not this hierarchical flattening has been achieved to any appreciable degree, however, participatory evaluation both requires and benefits greatly from direct lateral communications among field units and participating communities (Box 3.2 gives an example). Communication improvements can be achieved through careful design of representation schemes in which local units send delegates to central deliberations and those delegates then report back; but such a traditional "legislative" pattern is seldom enough and needs generally to be supplemented by more direct forms of lateral communication: delegations from one locality to another, document exchange and whatever other types of "networking" available technology and traditional circuits of social

communication make possible. The broader and stronger the base, the higher it is possible to build the edifice.

External communication

To this point we have been talking about communications up and down, or across, the different units and groups that populate the administrative structure of a program. But, as the notion of "stakeholders" discussed above implies, many more people and groups are in fact interested in the operations of the program - and may influence its success and impact - than the official administrators and beneficiaries. Participatory evaluation offers the opportunity to "broaden the net" by including these other actors in the loop and by developing new channels of communication with and among them.

Perhaps the simplest and most effective formula for doing so - though by no means the only one - is to open up the sessions planned for synthesis and interpretation of results at each level in order to include in them representatives of all key stakeholder groups, and then to use available media as a way of "airing" the results to a still wider public. The press conference, press release or public hearing can be one good way of including outside stakeholders in the evaluation "loop."

These three dimensions of communication about evaluation findings - the vertical, lateral and external ones - illustrate the challenge of "choreographing" a participatory evaluation so that optimum mileage is made of the information generated, and so that the different stakeholder groups begin to converge on a new consensus.

Making internal and external evaluation complementary: The concentric principle

Opting for participatory evaluation does *not* mean abandoning any plans for more detailed and technically sophisticated studies, or even for externally-executed evaluation exercises. It *does* mean engineering the evaluation design so that all these diverse efforts respond to a central set of questions, remain fundamentally comprehensible to all participants, and constitute in their eyes valuable additions to the common store of knowledge. That is not a requirement for bland uniformity or rigorous standardization. It is one, however, for a common core. And it entails developing what I would call a "concentric" pattern for the different studies and inquiries that make up the overall evaluation design.

"Concentricity" means simply that all portions of the evaluation design should respond to a common set of questions and fit (loosely, at least) within a common reference model, so that their results remain accessible and understandable to all participants. (Figure 3.2 represents the principle graphically.) In our NFE evaluation work in Mali, for example, the entire design was focused around five simple questions:

1. What have people learned?
2. How have they been able to use what they learned?
3. How satisfactory do the different groups involved find these results to be?
4. What aspects of the way in which courses are organized and taught, or the way in which the program is supported from the outside, explain the good or bad results obtained?
5. What ought to be done next year - in the local community and by the program sponsors - to improve things?"

Those five questions constituted virtually the entire evaluation at the village and literacy center level, though they were accompanied by some simple forms and self-testing procedures that enabled local participants to gather the data necessary to answer them. At the regional and national level, however, the scheme was more elaborate. It included analyses that could be done by aggregating and comparing the local-level results and by combining them with information about administrative operations and constraints available at the central level - yet it was based essentially on the exact same design. In addition, the national evaluation team, assisted by outside consultants, carried out some more detailed studies (case studies in selected communities, program impact measures, cost-benefit comparisons) that were beyond the immediate possibilities of local participants and agency staff. Once again, however, they basically provided information of use in responding to the same questions; and their methods and results were fed back into the participatory evaluation loop and explained to all participants. So a "concentric" design of evaluation questions and component surveys ensures that all information generated remains potentially accessible to participants at the local level and useful for them. It creates a medium of communication among stakeholders and a means for involving them in dialogue. Figure 3.2 illustrates the flows and relationships in question about a negotiated set of issues.

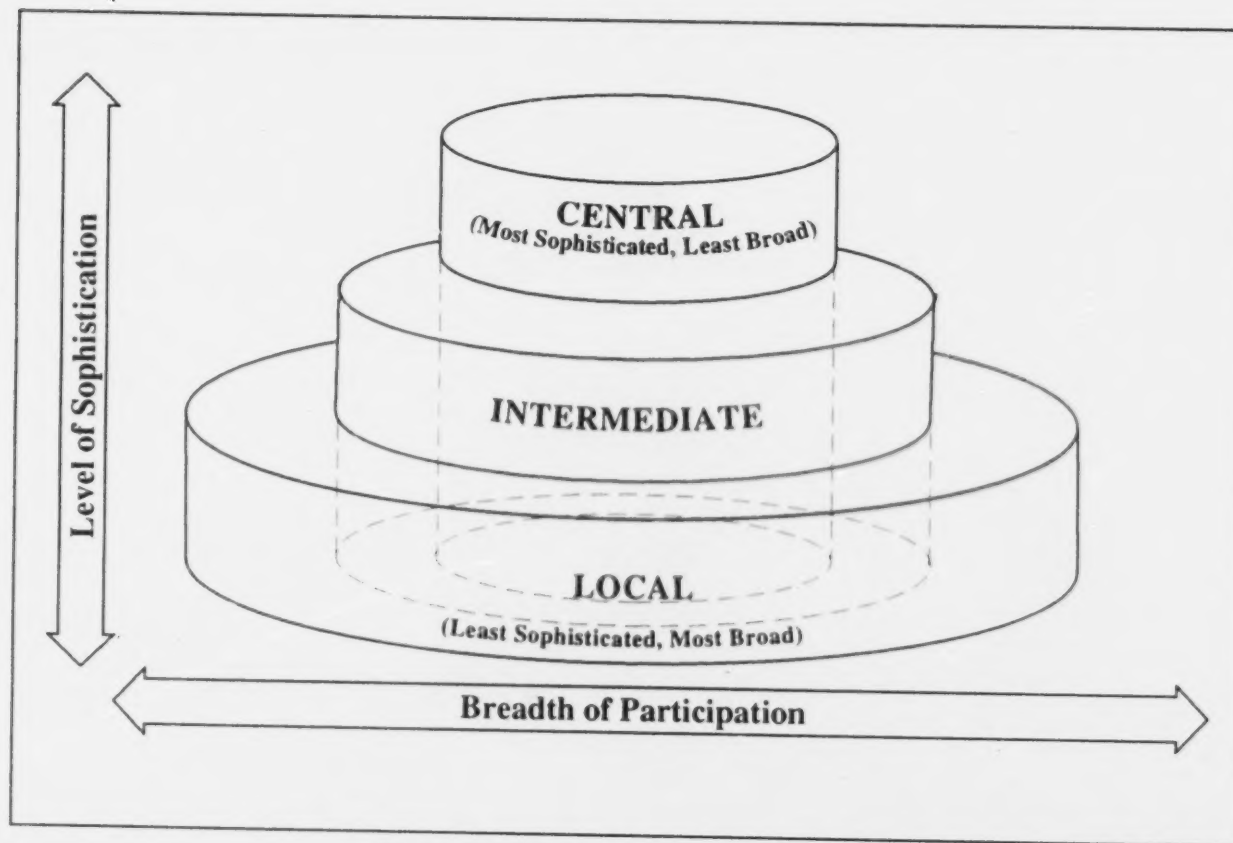


Fig. 3.2 Concentric design in participatory evaluation

Making sure it all leads somewhere

One of the traditional lessons of community development is quite applicable to participatory evaluation: make sure by every means possible (and legal) that the first activities undertaken yield visible results which those directly concerned consider positive - otherwise momentum is quickly lost and it is doubly difficult to start again. Though participatory evaluation offers a wide range of benefits, it comes with certain price tags as well. If there are no palpable results, participatory evaluation can in effect be done only once. Good luck trying to rouse interest and ownership for a second round among stakeholder groups (and particularly among intended program beneficiaries) if they feel that nothing came of their first efforts. It just won't fly.

"Results" does not necessarily mean that everything must change overnight nor that all problems must be entirely resolved. It does, however, mean some evident forward motion. Perhaps most of all and best of all, it would mean that participatory evaluation was followed by some degree of *participatory planning* - that is, an opportunity for beneficiaries and local staff to set (within the framework of common goals) their *own* objectives for the next phase of program activity, decide about the allocation of resources to accomplish these ends and find ways to begin remedying some of the problems that they themselves had brought to light. Though "participatory planning" may sound like a daunting initiative, and one well beyond the scope of this book, in practice it is often a natural outgrowth of the kind of evaluation exercises discussed in these pages and can be undertaken in incremental and quite simple stages. Moreover, the competence and experience acquired in participatory evaluation activities generally proves quite applicable to strategic planning exercises.

Box 3.3 gives two examples of participatory planning processes born out of participatory evaluation exercises. The first of these, taken from the Republic of Mali, is also illustrated in Figure 3.3 thereafter.

Evaluating the evaluation

It is generally a good counsel to practice what one preaches - or, in the words now popular in the Total Quality Management movement in business and industry, to "walk the walk" as well as "talk the talk." It is particularly advisable in the domain of participatory evaluation, because

Box 3.3. First steps in participatory planning: Two case examples

On the heels of a participatory evaluation of a functional literacy program in western Mali carried out in 1976 and 1977, the Direction Nationale de l'Alphabétisation et de la Linguistique Appliquée (DNAFLA) in tandem with the regional agricultural development company then sponsoring literacy interventions (OACV), asked participating villages to develop objectives and proposed methods for the following year's instructional program and post-literacy activities. Village councils started with the results of the evaluation just conducted and the recommendations made locally in the course of it for improving the program and, using much the same "participatory" methodology - that is, by soliciting the structured input of representatives of all local stakeholder groups - developed draft plans for priority activities to carry out over the year to come, including clear demands concerning the role that government support agencies should play in the process. The planning exercise had the further virtue of providing locally-defined objectives against which results for the subsequent year could be evaluated by stakeholders themselves in the next round of assessment.

In 1994 in the State of Florida (United States), the Division of Vocational Rehabilitation (DVR) of the state Department of Labor was confronted with the challenge of meeting a federal government mandate for greater client participation in the planning of vocational rehabilitation programs. The federal government was concerned that people with disabilities and their representatives themselves have an active hand in determining which services would be provided to whom, how and when. After deliberation, DVR staff and consultants decided that the most natural way of creating a participatory planning process would be to involve those client and stakeholder groups first in a participatory *evaluation* of the methods used by the Division to date. A statewide process of consultation and evaluation was therefore organized and furnished both the "conduits" and the pretext for involving the same stakeholder representatives in developing objectives and new programs for the agency's subsequent yearly plan.

the credibility of evaluators in proposing this innovation and this increased "transparency" in program operations to other staff and beneficiaries depends in part on their own demonstrated willingness to take the same medicine⁴.

An evaluation of the evaluation itself should therefore be an integral part of any participatory design, and an aptitude, even a zest, for self-criticism and frank feedback from others is certainly among the qualities

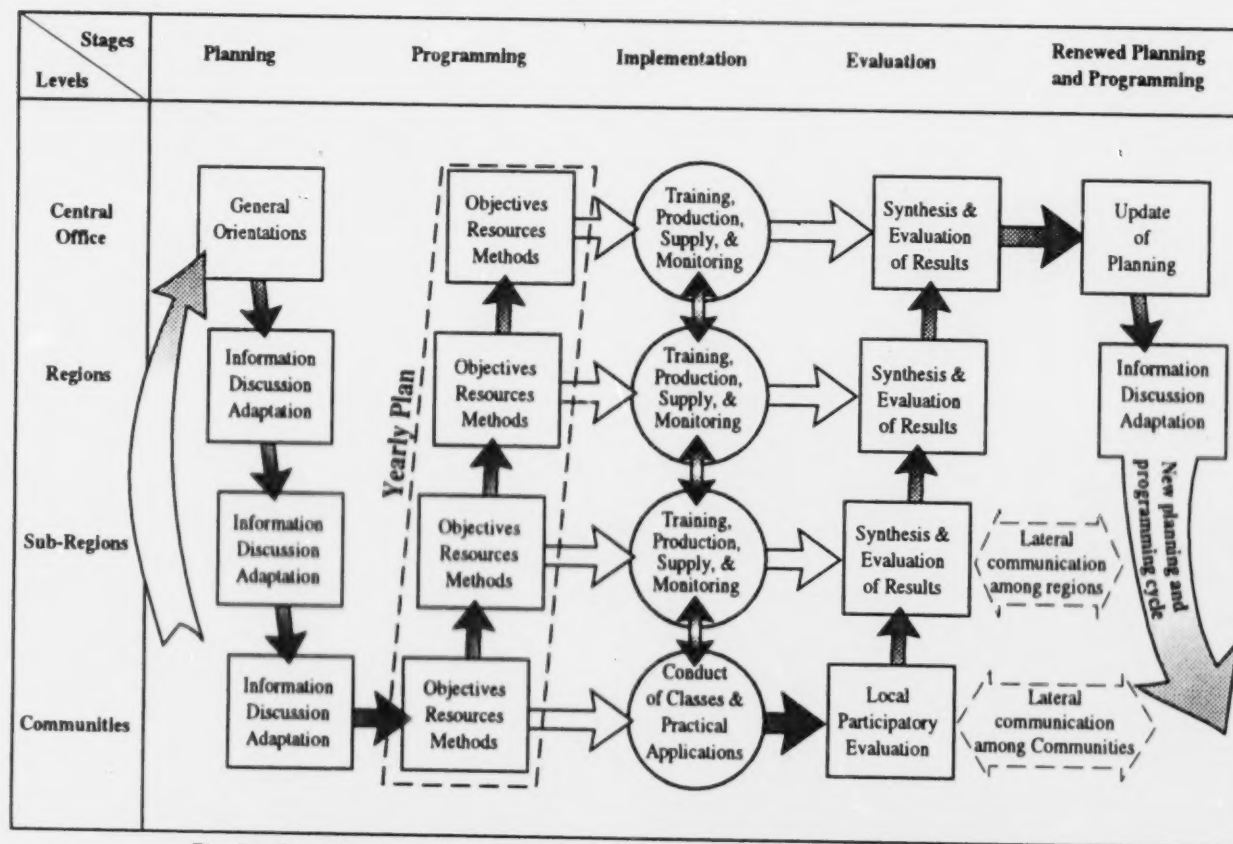


Fig. 3.3 One scheme for linking participatory planning and participatory evaluation in NFE programs.

and skills most needed by evaluators in these circumstances. This assessment of the evaluation itself, like any other, should include a "cost-benefit" dimension - that is, a comparison between the benefits of the evaluation and the resources invested in it; but one should remember that participatory efforts are always labor-intensive and generally manifest their impact more in the medium term than the short run.

Evaluating the evaluation and its methodology appears that much more critical when we view the whole undertaking as, in some sense, an exercise of research and development (R&D). If, as suggested at the beginning of this chapter, "disseminating the practice of evaluation itself" is one of the prime objectives of everything evaluators do, then it is evident both that the quality of an evaluation must be judged according to how well it has increased participants' competence in the assessment domain and that to do so involves a long-term teaching commitment. No one has foreknowledge of how to do this best in any given situation, any more than research and development staff in an organization know ahead of time exactly what product or service will best meet market demand or public need. The only thing to do is to develop and try out new approaches, assess them, and modify them on the basis of the lessons learned. Evaluation is just such a product or service: there are many lessons to be learned about how to disseminate it and improve its methodology, but they will not be learned unless we make a habit of assessing the assessors and their assessment.

Conclusions

Participatory evaluation is neither an afterthought nor a politically correct desideratum: it is an upfront principle. Several themes developed in this chapter will be interwoven throughout the rest of the book and are worth summarizing:

- "Participation" may be more or less genuine and thoroughgoing, but in its most stimulating and valuable forms it means enabling stakeholders and beneficiaries to take part in planning programs, allocating resources and the full range of decision-making regarding the activities in question as much as in their implementation.
- Evaluation is a doubly important arena for participation: not only does it offer ample opportunities for training and inducting stake-

holders into a variety of new skills and responsibilities, but in addition popular involvement in evaluation may constitute a vital "entering wedge" for developing higher levels of participation in the other domains and dimensions of the program in question. People who have gotten the kind of insight into program operations that a good evaluation can offer are much readier to play a role in planning, budgeting and supervising its conduct.

- Participation in evaluation, like substantial involvement in other dimensions of social activity, can be at the same time a highly *educative* exercise: one never learns quite so much as when one is called upon to assume new responsibilities and to take a larger view of the situation in which one is working.
- Participation and technical competence are natural complements, not opposites. The central theme of efforts to increase participation in technical affairs is that the fundamentals of the topic in question are not so hard to learn and can be formulated in a way that makes them accessible to the majority of interested parties.
- Participation has its own methods and techniques. Among these are the skills required to "choreograph" the participation, interaction and learning experiences of a variety stakeholder groups, as well as to analyze the technical issues in a "concentric manner" that locates, distinguishes and relates the basic notions accessible to all and the further refinements of greatest concern to technicians.
- Participatory evaluation that does not lead in some manner or form to opportunities for participatory planning through which those involved can provide for the application of the lessons learned in their assessment will likely be a short-lived affair.
- Participatory evaluators need to practice what they preach by providing for beneficiary assessment of the evaluation itself.

Happily, varieties of participatory methodology seem in this last decade of the 20th century to be becoming much more of a mainstay of evaluation practice, as well as a "hot topic" in professional and academic circles (cf. African Development Foundation 1995; Choudary & Tandon 1992; Fetterman et al. 1994; Guba & Lincoln 1989). But the proof of the pudding is in the eating, and the fate of participatory evaluation approaches will always largely depend on methods that practitioners develop in the field.

The subsequent chapters on evaluation technique and specific assessment methodologies will thus all be seen in part through this prism of participation. We will be asking how data quality concerns, indicator systems, sampling strategies, qualitative evaluation methods, learner self-assessment schemes and cost-benefit analysis procedures can all become vehicles for broadening participation in evaluation and increasing its learning value for those involved at the same time as they serve to increase its effectiveness and the validity of its results. An evaluator is a special kind of teacher or adult educator, one charged with helping the social groups allied in program implementation to upgrade their understanding of the task they have undertaken and to increase their capacity for perceiving and analyzing the results of educational intervention. Concern with the participation of these groups therefore underlies everything the evaluator does, and will underlie all topics broached in the rest of this book.

Notes

¹ One could say that the entire field of instructional systems design is organized on just such a premise concerning the "decomposability" of technical matters into hierarchies of learning.

² One corollary of this principle of "learnability" is worth noting: as a general rule, one can assume that if some subject appears too "technical" and difficult for a group of learners, it is because the person responsible for presenting it was incapable of making and communicating these kinds of distinctions, not because the audience was incapable of understanding them! Often the problem resides simply in the use of a technical vocabulary that, if the truth be known, the teacher him or herself may not fully understand. In such a case, better to stand by an old maxim of adult education: when in doubt, always use the *definition* of a term, not the term itself, as this helps both to facilitate listener understanding and to ensure that the speaker also has a good grasp on things!

³ In North American business and administrative circles, "management by the objectives" has in fact gotten a rather bad name of late. The approach has been characterized as a way of imposing central management's priorities on organizational members or employees and requiring them to accomplish only what the boss judges important. Though the delegation of responsibility that this method of administration implies and entails can really liberate creative energies, it is clear - particularly from a "participatory" point of view - that

its potential cannot be realized unless the "objectives" in question are *jointly-defined* and mutually valued ones.

⁴ Such a willingness to be hoist with one's own petard is not always a leading characteristic of evaluation and research personnel, unfortunately. As the Sahelian proverb puts it, "The blood-letter never likes having his blood let"!

CHAPTER 4

SYSTEMS THINKING AND PROGRAM THEORY

Even if they deliver relatively simple services to their intended beneficiaries, nonformal education programs are complex undertakings. Over and beyond the obvious details of teaching and learning and course organization that quite properly occupy front stage, the operation of these programs depends on a number of factors "upstream and downstream" from the instructional domain. They are connected directly or indirectly to a series of other institutions in their social environment, and they are influenced by a welter of cultural customs and human needs. Which of these other factors should one include in the focus when evaluating such programs or helping their staff and beneficiaries to get a better and broader grasp of what is happening? If too much is left out of the picture - if, for example, attention is focused strictly on the learning transaction without taking into account any of the rest of the environment - then one risks misinterpreting results, rather like an extension agent trying to understand why harvests are poor simply by looking at the way in which seeds are planted. On the other hand, if one tries to pay attention to everything at once, especially without some principle of organization, then the picture becomes so jumbled that it is hard to discern any pattern or to know where to stop.

Many schemes could be used for organizing evaluation data about a program. In this chapter, one will be proposed that is not necessarily superior to the rest but that does provide a sort of "coat rack" for arranging and analyzing evaluative information on NFE programs as well as a check-list for making sure that no essential elements have been forgotten. It is derived from the tradition of systems analysis, a loosely defined set of methods for examining the many different actors and forces that influence the conduct and outcomes of a social undertaking. It furnishes a "common thread" around which data analysis and interpretation exercises can be organized and a "common language" through which the many parties involved in a participatory evaluation can exchange information and viewpoints.

Let's begin by reminding ourselves, however, that none of these schemes or models has any eternal or transcendent value. They are

simply instruments that we use to facilitate our reflection. The habit of devising and revising representations of a program and models for evaluating the program is much more important than any given model or representation itself. One might say, tongue only partially in cheek, that such models can be highly useful as long as they remain both "plastic" and "biodegradable" - that is, as long as one is willing to modify them and ready to throw them away periodically for recycling. They make excellent servants but very poor masters.

Systems modeling of NFE programs

The model proposed here is generally grounded in the tradition of systems analysis. A methodology for assessing the performance of complex operations, systems analysis was devised to help meet production needs during the Second World War and then used for many other purposes. It is based on a fundamental distinction among five elements of any activity.

1. the environmental "context" in which it takes place;
2. the "inputs" or resources (material, human, financial and so forth) that it requires in order to take place;
3. the "process" or methods by which these inputs are then used, combined and transformed into desired results, be they services or goods;
4. the "outputs" created in this way - that is, the services and/or goods that result from this effort, including its unanticipated by-products; and
5. the longer term "outcomes" which these results provoke - namely, the eventual impact of the activity as a whole on its environment.

This process is, in a sense, circular. The activity or program in question produces some sort of impact on the environment, modifying the initial conditions of the system. That may in turn change the resource base of the activity and therefore its inputs, as well as the way in which it is carried out. A graphic image of this scheme is presented in Figure 4.1. Evaluators will quickly recognize the close relationship of this diagram to Stufflebeam's CIPP model (later modified to "CIPOO" with the differentiation of "products" into "outputs" and "outcomes"); but it has

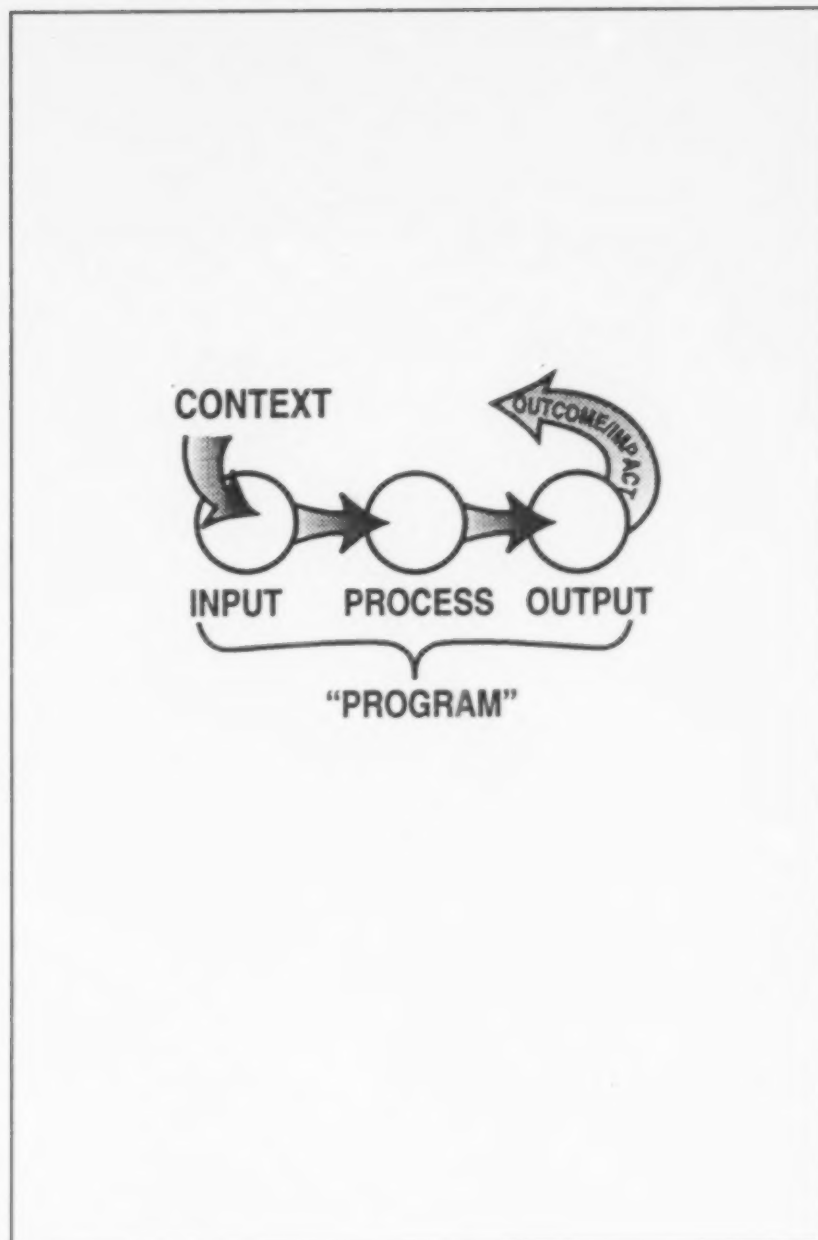


Fig. 4.1 Adaptation of systems model for evaluation purposes.

been modified and elaborated here in particular ways (Stufflebeam 1971).

How is a scheme like this used for evaluation purposes? Social programs can be viewed as small-scale systems set up within an existing environment for more-or-less predefined purposes - generally to bring about certain outcomes judged socially desirable by those in charge. Desired outcomes represent the goals or longer-term objectives of the program - what its sponsors would like to change in the environment - and therefore they correspond to somebody's assessment of what is deficient in that context at present. Outputs represent the immediate results that one would need to produce in order to have a chance of creating that longer term impact: if a public health agency hopes to increase the use of contraceptive methods among rural women and men, it will need as a preliminary result first to increase both awareness of these methods and availability of the necessary materials. Inputs represent the sorts of resources the program will require in order to produce those results (e.g., educational materials, competent public health educators, etc.), whereas processes describe the program methods (for example, type of curriculum, instructional techniques and social marketing strategy) that will be required to achieve such ends. The necessary inputs and the methodologies to be used will have to come from somewhere in the "context" - that is, the surrounding environment, whether nearby (village, town) or farther away (national capital, foreign organizations).

A program can therefore be thought of as essentially an assemblage of these elements:

- *Context:* Some understanding of the environment and an identification of particular problems within it that need to be remedied, or particular gaps between the present situation and a more desirable one.
- *Inputs:* Specification of the resources required to resolve the problems that have been highlighted, including the time necessary to do so.
- *Process:* Definition of the intervention methods or actual programs that should be undertaken with these resources in order to remedy the situation and produce durable change.
- *Outputs:* Specification of the immediate results that this activity is likely to produce.

- *Outcomes:* Projection of the longer term impacts on the participants and their environment to which these short-term changes can lead, and the conditions under which such things could happen. Typically, these longer-term impacts correspond to the underlying goals of the program and amount to "closing the gaps" - or resolving the problems - defined above in the initial analysis of the context.

These five elements in fact mimic the form in which most project proposals for funding by national and international sources are prepared and submitted. The United States Agency for International Development used for years as a template for project design a "logframe matrix" which essentially reproduces, with some modifications, this same scheme and which is pictured further on in Box 4.3. There is even a branch of recent thought about evaluation called "program theory" which basically maintains that every social service program embodies a theory about how given social problems are remedied and behaviors modified that can be portrayed in this way - and we shall return to it below.

Evaluation dimensions and program domains

But first, let's try to put a little more life and specificity into the portrait, since the language used above remains highly generic and seems to have little connection with the specific characteristics of NFE programs. Just what resources are needed in such programs? Just what methods are used? What kinds of short- and longer-term results are targeted? How do these interact with each other? And what other factors in the environment of the programs have the most important influence on them? Figure 4.2 portrays an adaptation of the systems model developed for the evaluation of literacy and adult education programs in Burkina Faso, West Africa. Several things are worth noting about this scheme:

- First, when the notions of input, processes and outputs are projected onto the screen of an entire activity like the literacy programs in question, an interesting distinction appears. In macro-analytic terms, the "input" domain corresponds roughly to the administrative realm of the system, which lies "upstream," so to speak, of the instructional activity itself. In fact, the primary responsibility of administrative services in a program is to obtain, develop and provide the resources that the program needs in order to be implemented in

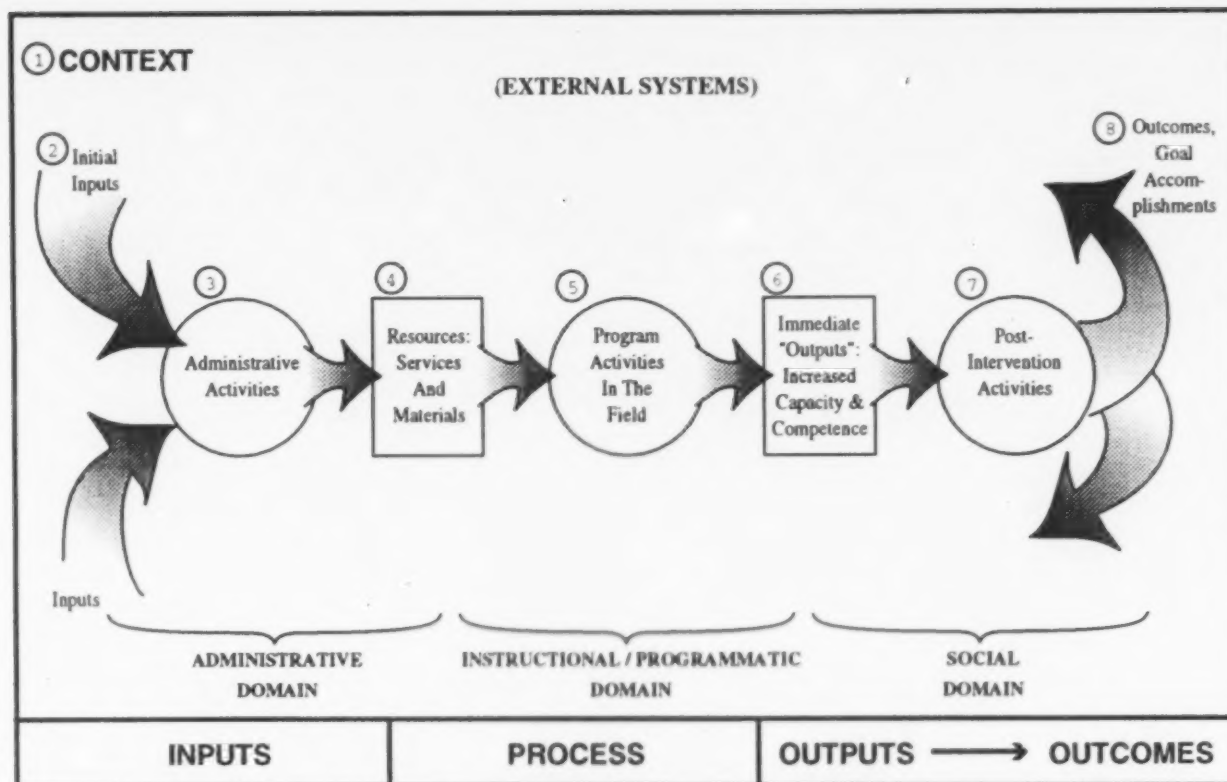


Fig. 4.2 System model applied to instructional programs

the field. It makes sense, therefore, that program administration figures on the input side of the diagram.

- The "process" domain, on the other hand, corresponds in overall terms to the instructional realm of activity - the locus of the actual teaching and learning which is the main focus of the program. The instructional subsystem uses the inputs provided by program administration and other sources in the environment to enable teaching and learning to take place. The immediate outputs of this process and of the instructional domain are obviously constituted by the learning results of the activity: i.e., the knowledge, skills and attitudes acquired by participants.
- The output/outcome domain coincides with what one might call the "post-training" or "post-literacy" realm of activity: all the undertakings through which new literates or former trainees manage to use and apply their newly acquired knowledge. The results of instruction and, in some cases, the resources and support provided by program administration or other stakeholders constitute inputs to this domain and help to make post-training initiatives possible.
- Finally, the context or environment obviously surrounds the program on all sides. How inclusive one decides to be in portraying this context or in the "level of resolution" adopted is a matter of choice that will depend partly on the level of resolution used for the system itself: Is this scheme being used to represent a national program (as in the Burkina Faso case), a regional one, or a strictly local undertaking?

There is another very important thing to note with respect to the entire diagram. One critical dimension of the program and its context does not appear very fully here, because in effect it is stacked "behind" the visible diagram - namely the fact that the resources and methods devoted to the program, and the results that it obtains, are actually superimposed on similar existing functions within the environment. In the case of the Burkina Faso literacy program, there were both a succession of preceding programs that had left traces in each village and locality and, in a deeper but still contemporary sense, traditional patterns of training and continuing education which - even if they did not much resemble the contemporary government-organized or NGO-organized adult literacy efforts - did serve some of the same functions.

This is indeed a critical point. One of the prime factors of success or failure in programs is the degree to which they manage to build on the best in existing habits, institutions and traditions and to develop capacities already present, or latent, in the environment. If the program is simply "imposed" from the outside, as this diagram might lead one to suspect, it has little chance of provoking durable change or improvements, because long-term results are by definition the result of incorporation into permanent institutions (as well as creation of new ones). It needs either to reinforce and develop existing functions within the environment or constitute a viable "graft" onto them. For the diagram adequately to represent this dimension of reality, we would need a 3-D image in which the "shadow" or backdrop of existing habits and institutions was clearly visible behind the program model.

The "Russian doll" effect

Next one can note that the input-process-output distinction is quite relative, a matter of scale and focus, because the "system" portrayed in this way is in fact composed of a series of subsystems, each of which has its own inputs, processes and outputs. These systems and subsystems can be seen as "nesting" within each other rather like Russian dolls. Thus the administrative realm is in fact a system of its own, with inputs (funds, people, and equipment), methods (the procedures of administrative organization and agency function by which these resources are put to use) and outputs (the services and materials delivered to NFE centers in the field) which are particular to it.

Likewise the instructional subsystem can be considered on its own terms: it uses inputs that are principally, in fact, the "products" or outputs of the administrative domain (teaching competencies, learning materials, supervisory services and so forth); its own processes are represented by the instructional and organizational methods necessary to facilitate learning; and its own outputs are the various competencies acquired by participants.

Finally, post-training activities constitute an additional domain of concern "downstream" of the instructional program. Here the inputs comprise, on the one hand, the competencies acquired by program participants and, on the other, the various other resources necessary for them to put this knowledge to personally or socially productive use (books and newspapers, jobs, resources with which to start new busi-

nesses, etc.). The processes concern the detail of these "post-literacy" activities, whereas the outputs concern their results: changes in the knowledge levels, production levels and well-being of people and their communities. These are, at the same time, the indicators of the overall "outcomes" of the training program and constitute the impact that it produces on its environment. Table 4.1 summarizes these varied dimensions of an NFE program and offers a visual presentation of the overall model.

Yet things can be broken down even further within each of these large domains. The administrative domain provides a good example: teacher training, instructional material design and production, and field office supervision could each represent a system unto itself that one might analyze - and evaluate - in terms of its own resources, methods and results ... and the story goes on and on. As the childhood rhyme puts it, "Big fleas have little fleas upon their backs to bite 'em; and little fleas have smaller fleas, and so *ad infinitum*"! In Figure 4.3, a more detailed systems model of an NFE program - originally composed as a support for analysis of literacy programs in the Republic of Mali - is portrayed. Here the "sub-systems" in each domain appear more clearly and the flow of activities among them is graphically indicated.

The point once again lies more in the method of analysis and the light that it can throw on program operations than in any particular configuration or "level of resolution" specified rigidly *a priori*. What is more, it should be obvious that the diagram can be modified however one sees fit in order to represent more accurately or usefully the particular program with which one is concerned. We shall, in any case, return to this sort of systems portrait of NFE program operations on repeated occasions in the rest of the book, because it offers a very useful way of depicting the different forces, factors and stakeholder groups that jointly determine program results.

This type of input-output scheme can be (and has been) taxed with taking too "productivist" an orientation to education and training, imposing a "factory model" on the act of learning, and/or reducing the results of programs to their economic and quantifiable dimensions. I nonetheless find it helpful and prefer to think of it (and use it) as a more generic tool than that - as an instrument once again to facilitate reflection. "Output," from this point of view, is simply a conceptual metaphor for results of whatever nature, emphasizing that they do not fall from the sky but are to a considerable degree a consequence of what is and isn't

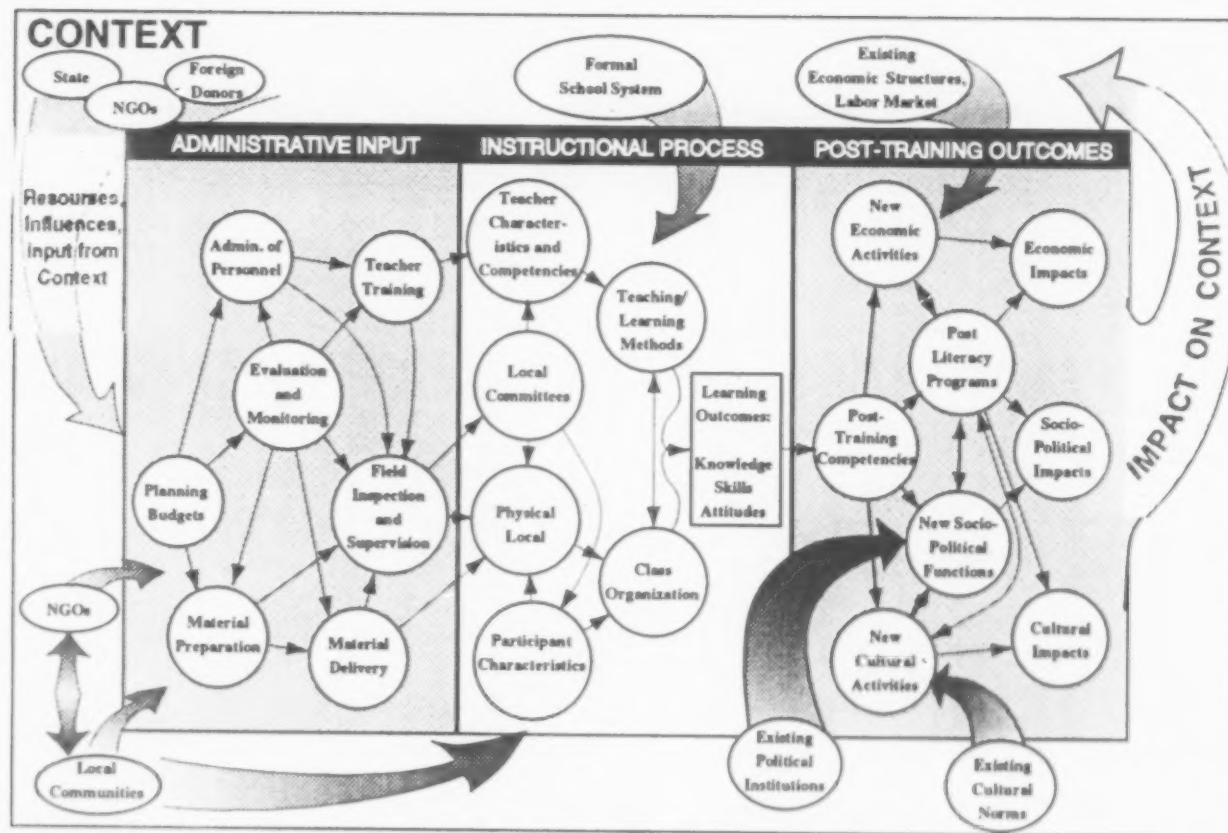


Fig. 4.3 Systems representation of an NI-BE program

Table 4.1. Tabular presentation of the program or intervention strategy of an NFBE program

	Administrative Domain (Program organization)	Instructional Domain (Teaching and learning)	Socio-Economic Domain (Post-training activities)
Assumed Context	e.g. norms of administrative behavior	e.g. methods of other current and past literacy efforts	e.g. level of development of target localities.
Intended Input (Resources)	Funds, personnel, material, political support and conceptual resources to be devoted to program. (bequeathed inputs)	Instructional services, learning materials, and logistic support to be provided to centers. (produced inputs)	Resources necessary to make training applications possible, other than the knowledge and competencies of former trainees
Process and Methods	Procedures and methods by which administrative entities produce the services & material needed.	Teaching and learning strategies used in the instructional center or activity.	Methods for organization and conduct of post-training and application activities.
Immediate Output (Results)	Quality, quantity and timeliness of services and material actually delivered in the field.	New knowledge, skills and attitudes acquired by trainees.	Short-term improvements in personal and social welfare resulting from above.
Longer-Term Outcomes (Impact)	Contribution of above to smooth functioning of instructional and socio-economic domains.	Contribution of these to post-training success & improved functioning in socio-economic domain.	Long-term improvements in personal/social welfare resulting from training & its applications.

66

Systems thinking and program theory

85

BEST COPY AVAILABLE

86

done in connection with the program. Changes in attitudes, self-discovery, religious conversion, political consciousness-raising and development of social solidarity are just as legitimate as results - and should be as easily fitted into this scheme - as are test scores, job placements, or income changes.¹ Similarly, "input," in this context, is just a term for all the resources that we invest in an endeavor in the broadest sense (i.e., including people's time and commitment, political allegiances, and potential environmental sacrifices as well as funds and material). The principal virtue of this kind of model is simply to furnish a language for collective reflection on what is being accomplished with the resources at a program's disposal and how it is being accomplished. It is ultimately therefore a tool for accountability.

Program theory

Most NFE programs funded by outside donors - whether local or national governments, private foundations, business interests or international agencies - have some sort of explicit program design. They have stated objectives and methods plus a budget indicating the resources that they require and how these will be used. They are expected to be accountable in some manner to their sponsors and/or beneficiaries for the use of these resources and for the results achieved. (In fact, the practice of evaluation developed initially in large measure as an accountability requirement of outside donors and has only gradually transformed itself into a tool for program improvement by national and local parties.) Evaluation is therefore often conducted with explicit reference to the design and intentions of the program: were these carried out, were the anticipated results obtained, and if not, why not? Moreover, those designs and intentions constitute in some sense a theory about how social interactions work in the domain of intervention and how cause and effect are linked: in short, a "program theory."

As Michael Scriven first pointed out over twenty years ago, however, this focus on official goals and objectives is not entirely adequate as an agenda for evaluation and risks blinding evaluators to many of the actual effects of programs (1973: 173). First, official goal statements and even methodologies are often more political statements than realistic descriptions of what a program intends to achieve. There are all kinds of political reasons that push program designers towards remaining very

vague and general in setting initial objectives, among them the need to gain the consent of opposing interest groups without appearing to favor any one, and adherence to the time-honored bureaucratic principle that if you don't state clearly what you intend to do, no one can reproach you later for not having done it. In addition, even supposing original program designs are entirely earnest and honest, they almost always change form over time as they are implemented - perhaps because what was proposed turns out not to be feasible, or simply because it is interpreted differently by the field staff responsible for implementing than it was by the program designers. Mutations in program purpose and methodology are seldom recorded as clearly and explicitly as the original design. And finally, there are a good number of programs, particularly those of local origin, that may have grown up without any explicit initial design at all. (See Box 4.1 for an example.)

To solve this dilemma, Scriven suggested the practice of what he called "goal-free evaluation." By that, he meant an assessment of present practice and all discernible results not based on or limited to what official statements said the program was supposed to be doing. This approach has the considerable virtue of opening the evaluator's eyes to many more things than he or she might otherwise examine, and particularly to the kinds of unexpected interactions between program and context that often offer the most useful insights into desirable future directions.

Box 4.1 Programs that "just grew"

The APOV community education and development program in the Bela Horizonte province of Brazil was born from the determination of its founder, a Catholic laywoman from Vicosá, to do something for the poor in the neighboring satellite community of Nova Vicosá. Methods and activities were added bit by bit as funding became available and needs became evident, and the program presently includes eight major components, reaching *in toto* more than 5000 people. Yet if you ask for a program design document, you will not find one. Activities were developed and added to the initial undertaking "by the leading of the Holy Spirit" and as resources and personnel permitted. The program has nonetheless its own internal logic and an implicit set of objectives. The evaluator working with APOV spent some time reconstituting these with staff and participants before trying to investigate or assess the current operations of the program (Gareau 1995).

At the same time it remains true that, if evaluation can be to a certain extent "goal-free" at the input side - that is, with respect to selection of topics to investigate and types of objective accomplishment to assess - it remains necessarily more goal-oriented on the output side. In other words, the purpose of evaluation is in good part to enlighten social discourse and decision-making about the area of concern defined by the program being evaluated, and to do that by deriving from an examination of the program's procedures and results some improved (and communicable) understanding about how causes and effects are related in this area, what kinds of interventions produce the best results under different sets of circumstances, and how existing institutions and groups can best be helped to devise or adapt strategies that will durably alleviate the problems in question.

Evaluation is therefore itself a generally goal-oriented activity, and its goal is to derive from the experience of the program usable knowledge and "workable theories" about how to remedy the social problems that it addresses. As Dennis Rondinelli has pointed out, projects can in many respects be regarded as "policy experiments," and experiments are manifestly designed to be evaluated (Rondinelli 1984).

This is what is meant by the notion of "program theory" - or perhaps, more accurately, "program theories." Whether or not a program has been designed in an explicitly documented manner - with detailed specification of its objectives, methods, budget and so forth - it implicitly functions according to some sort of organic model, or amalgam of models, that determines what resources shall be obtained for the undertaking, to what uses they shall be allocated and how, what specific activities shall take place in the field, and what sorts of results or "payoffs" should be anticipated by the different stakeholders involved. This implicit theory or intervention model is, moreover, basically grafted or superimposed onto existing social habits and functions and may either reinforce them, modify them, or conflict with them.

It is the evaluator's task to help stakeholders articulate the implicit "program theories" or intervention models on which the action is currently based and to draw from their experience the sort of "usable knowledge and workable theories" that will both help the program function better and provide insight to others who might wish to replicate it or to undertake parallel efforts in their own contexts. Such an effort is particularly critical in the case of a participatory evaluation, and in that of participatory adult education programs in general, because these pro-

grams greatly need to be based on local knowledge and adapted to the particular potentials and needs of each locality. That is generally the condition *sine qua non* for promoting "ownership" of the activity by its intended beneficiaries and for ensuring thereby its sustainability. Given that most programs are designed at the outset in large part by outside parties - whether government technicians, foundation officials, NGO personnel, or international agency staff - evaluation becomes a critical means for enabling local stakeholders to articulate their priority objectives and their own version of program strategy, both by analyzing the strengths and weaknesses of whatever portions of official methodology have been effectively implemented and by documenting the dimensions of variance that have emerged through local adaptation of this model. Evaluation becomes one prime means of developing a genuinely "indigenous" model from a partly or entirely imported one.

What approaches can be of use to evaluators in this task of articulating, testing and "indigenizing" program theory? Three will be highlighted in the rest of this chapter: methods for reconstituting program theories, techniques for drawing up and analyzing the "balance sheet" of program results, and ways to deduce new potentials from the comparison of program theories and program results.

Reconstituting program theory

It was suggested above that every program operates to some extent according to an implicit set of theories or models that indicate what one does and why. In fact, programs with explicit intervention models may in fact operate not in conformity with these, but according to subsequent unpublished versions that have resulted from the many compromises necessary to implement the activity in the field, or according to implicit arrangements that are not meant to be public knowledge. Worse yet (for evaluators), programs may operate according to several different schemes at once, and/or different stakeholder groups may have rather divergent views of what is in fact the real program model.

How to make sense of this tangle of versions and disagreements? One should of course start with whatever written documentation exists, whether official or unofficial. Even if the models described in writing bear little resemblance to what is done in the field, they at least demonstrate what someone in power (or equipped with a word processor and

printer!) thought it necessary to say in order to justify the project, and that itself is useful evidence.

Next evaluators will want to interview representatives of key stakeholder groups (central staff, regional staff, intended beneficiaries, perhaps local authorities, etc.) about what they see as the essential methodology of the activity, and/or what successive versions of this methodology have been developed. Not everyone can articulate a synthetic view of these matters at the drop of a hat, and some who can may be inventing things; so obtaining such insight may require both prompting and editing. One useful technique for priming the pump and stimulating reflection on the program theory or model is to do a group exercise devoted to reconstituting it.

The exercise usually proceeds in reverse order. The first questions concern the essential long-range objectives of the activity. What impact is it supposed to produce on the environment? Next one asks what conditions must obtain in order for these impacts to be produced - in other words, what are the short-term objectives? The question that follows logically is how these short-term objectives are to be attained. What is the methodology to be used and what are the resources required? Here one will want to go into some detail, spelling out just who is supposed to do what, and with what support. At each level of the foregoing discussion, the facilitator or evaluator can mark on a chalk board or flipchart the different objectives and activities defined and link them to each other with arrows and lines, rather like the scheme portrayed in Figure 4.3. When one has gotten back to the responsibilities of central agencies and the support that they receive from the government and/or community and donor groups, the picture is probably nearly complete, or at least quite workable. Copies of the entire scheme can then be made so that everyone can look it over and correct it.

The process need not be as belabored as portrayed here. It can consist of no more than a collective interview with stakeholder representatives (or serial interviews with each of them), devoted to talking about the objectives, methods and resources of the program, and the ways in which they have changed over time.² Whatever the case, in my own experience, two things almost always result: First, the experience of reconstituting in this manner the program theory underlying their daily actions turns out to be immensely stimulating for the staff and stakeholders who participate and tends to provoke a whole new series of analyses and reflections on their part. Second, one generally ends up with several

different, if overlapping, versions of program methodology, because each stakeholder group sees the question a little differently. And that itself is a result of very great interest.

Successfully "teasing out" program theory from the perspectives of different staff and stakeholders requires, among other things, good interviewing technique and decent facilitation skills. One needs to be adept at interesting people in talking about their own work, analyzing it, and contrasting their varying viewpoints - without provoking unnecessary conflict. The evaluator must not appear either to wish to dictate her or his particular understanding of the program or to judge prematurely what stakeholders have to say.

In addition, "rapid turnaround" of the results of these sessions is usually highly desirable. In other words, evaluators who are able in relatively short order to capture the essence of what their interviewees

*Box 4.2. Building program theory:
The Educational Restructuring (BOETP) Project in Thailand*

Starting in the early 1990s in Thailand, the Thai government and the United Nations Development Program (UNDP) sponsored an initiative designed to decentralize responsibility for a number of educational improvement activities - both formal and nonformal - to the regional and local levels and to enable local school boards and teaching personnel to devise their own projects and submit them for funding. The activity required as a parallel initiative organized efforts to train local personnel in project design and implementation methods, including identification and adaptation of the "program theory" proposed by BOETP's sponsors.

In training sessions organized in Thailand and abroad, regional officials practiced detailing program theory by analyzing the causes to which high rates of school dropout in selected communities could be attributed. They specified a matrix of contributing factors, identified those that local initiatives could effectively address and remedy, brainstormed program elements that might resolve such problems, and assessed the methods and resources that would be required to do so. Principal among the contributing causes identified were inadaptation of curricula to local development needs, insensitivity of teachers to cultural diversity issues, and the need of impoverished families to retain children for productive labor. Each of these had in turn other sources that were identified. Only some could be addressed by project interventions. But the exercise greatly clarified participants' thinking about project strategy.

have revealed, to blend it into a synthetic presentation of program theory (and one that cites their actual words and/or those of relevant agency documents to the maximum extent possible) and to relay it back to them will probably have the greatest success in generating a consensus about the underlying strategy of the program and a workable reference model for subsequent evaluation work. It's a little like the process that the Chinese revolutionary leader, Mao Tse Tung, used to call "from the people, to the people."³

Box 4.2 provides an example of this process at work in the development of a new educational initiative.

It should now be possible to put together a sort of synthetic portrait of the methodology of the program (or several of them!), a scheme which uses the distinctions among administrative, instructional and post-training domains defined above and, within each, specifies briefly the resources needed, the principal methods proposed, and the results anticipated. The resulting portrait of the program might have roughly the form pictured in Table 4.1 above. (An alternate one, mentioned briefly already and prominently used by the United States Agency for International Development in the planning of projects and termed the logical framework - or "logframe" - matrix, is illustrated in Box 4.3 hereafter.)

More than one portrait may be required to give expression to the different versions of program theory currently in vogue, or that have been sequentially adopted over time. Whether one goes to these lengths in reconstituting the program model, or simply adopts a single "middle-of-the-road" version for reference, is a matter of choice. If sponsors or stakeholders count on the evaluation as an opportunity to rethink program design or to reconcile differing viewpoints about it, then it will probably be worth the evaluators' time to "explicate" these variant versions of the strategy in some detail.

This kind of synthetic presentation lends itself immediately to some useful analyses. To begin with, the articulation of different versions of the underlying program strategy may provoke quite a stir among program stakeholders and lead to some interesting discussion and perhaps a reconciliation of differences. Next, the simple fact of laying out, side by side, what the program purports to do in the administrative, instructional and post-training domains can be quite a revelation to those involved - who may never have pictured the situation to themselves in quite this way. At the same time, the exercise may bring to light certain incoherencies in the overall strategy. One is led, for example, to ask, "Is it

Box 4.3. USAID's logframe matrix

For years, the United States Agency for International Development prescribed a framework for design of the projects it funded that was popularly referred to as the "logframe matrix." Its purpose was to lead project designers through the series of steps required in order to define clearly just what the new initiative was supposed to accomplish, how it would do so, and how one would know whether or not it had: in short, a concrete embodiment of "program theory." The logframe matrix had the following general form:

	End-of-Project Conditions	Assumptions	Objectively- Verifiable Indicators	Means of Verification
Input				
Process				
Output				
Outcomes				

"End of project conditions" referred to the situation that should obtain in each of the domains once the intervention was complete: new resources installed, new methods disseminated, specific results achieved, specific impacts ensured. "Assumptions" were the preconditions or the factors beyond project control (whether material, political, social or psychological) that needed to exist for the intervention to produce with success the end-of-project conditions. "Objectively verifiable indicators" and "means of verification" both concerned the evaluation mechanisms put in place for determining whether or not goals in each realm had been achieved.

This scheme, while rigid and schematic in many ways, did also provide a very useful tool to evaluators in subsequently assessing the successes and failures of project implementation. On the basis of such specifications, it was possible not only to investigate the degree of accomplishment of intended results, but also to investigate stakeholder perceptions or this scheme and begin understanding the factors that explained *why* and *how* results had or had not been produced.

reasonable to expect *those* outcomes or results with *these* strategies and resources? Do the methods prescribed fit with each other, and do they take sufficient account of past experience in the field?" And so forth. This effort can be a prime instance of what Argyris and Schon call "organizational learning" (1978). In short, the exercise of reconstituting the program model becomes itself an early phase of participatory evaluation - before ever gathering the first shred of additional data. Of course, what really happens is that this experience causes stakeholders to bring to mind and to critical examination the experiential data that they already possess, and to use it in a first round of evaluation of the program - a fine example of drawing on existing resources! As qualitative researchers say, the ultimate instrument in any inquiry is the human mind.

Drawing up the balance sheet

In a sense, most of the rest of the evaluation now consists of "filling in the blanks," though the process is a good deal less mechanical than that phrase implies. It involves drawing up, so to speak, another two tables, similar in form to the one in Table 4.1, but consisting on the one hand of what has actually happened in the field in each of these realms under the aegis of the program, and, on the other, of the improvements in these different dimensions of program practice that could be proposed. The first of these contrasting tables - the one based on observation of actual practice - contains what are usually considered the data of an evaluation and responds to questions like the following: What resources have actually been received and used in each domain? What methods and activities are really being implemented? What results can one discover and document? This table constitutes, so to speak, the current "balance sheet" of the program. In fact, by comparing results obtained to resources invested within the table itself, one obtains something loosely resembling a "cost-benefit" analysis, without all the mathematical trappings. (For a few of the mathematical trappings and underlying justification of cost-benefit analysis itself, see Chapter Thirteen.)

The second table of "potential improvements" is based on the analyses and commentaries of different stakeholder representatives with regard to the "balance sheet" just composed as well as the perspectives of the evaluators. Both parties may also draw upon the example of other

activities transpiring in the environment of the project that suggest new approaches or better adaptations of existing skills and customs. Once again, one may come up here with several different versions of "potential improvements," depending on which stakeholder group is concerned. What's more, the evaluators might come to conclusions different from those reached by the interviewees. So be it: it is probably from the comparison and "triangulation" of these different points of view that the best recommendations for program improvement will result.

Comparing program theory and program results

There are a multitude of possible axes of analysis and comparison both within and among the different tables constituted in this way. If results actually achieved do not coincide with objectives, might this be because the resources actually used were not the same as those promised, or because the methodology implemented in the field was at variance with the one designed. And if so, why? If results fall short of expectations despite the fact that adequate resources were invested in the activity and the prescribed methods used, does this demonstrate that the program methodology itself is poorly adapted? Are there better alternatives out there in the field that one might use as a basis for program reform or revision? What do different stakeholder groups feel about this?²

Many of these contrast are summarized in the flow diagram of evaluation pictured in Figure 4.4. Here one sees that programs generally arise out of some variety of needs assessment, whether formal or informal, in which information about the current situation in some area of concern is compared with the values or standards of relevant stakeholder groups - that is, with an image of the situation that they would prefer. Of course, different stakeholder groups may have both different values or standards (portrayed in Figure 4.4 by the "shadow" circles IA and IB) and different views of the current situation; and program design is generally a product of compromise among them.

In any case, out of this comparison is born the will to make changes and a strategy for beginning to modify the present state of affairs and transform it into the desired one. This strategy serves as the basis for the development of the program model. The model is implemented over time with more or less fidelity and success and may, as was intended, exert important influence on changing the situation initially targeted, though that situation is doubtless also continuing to change under the

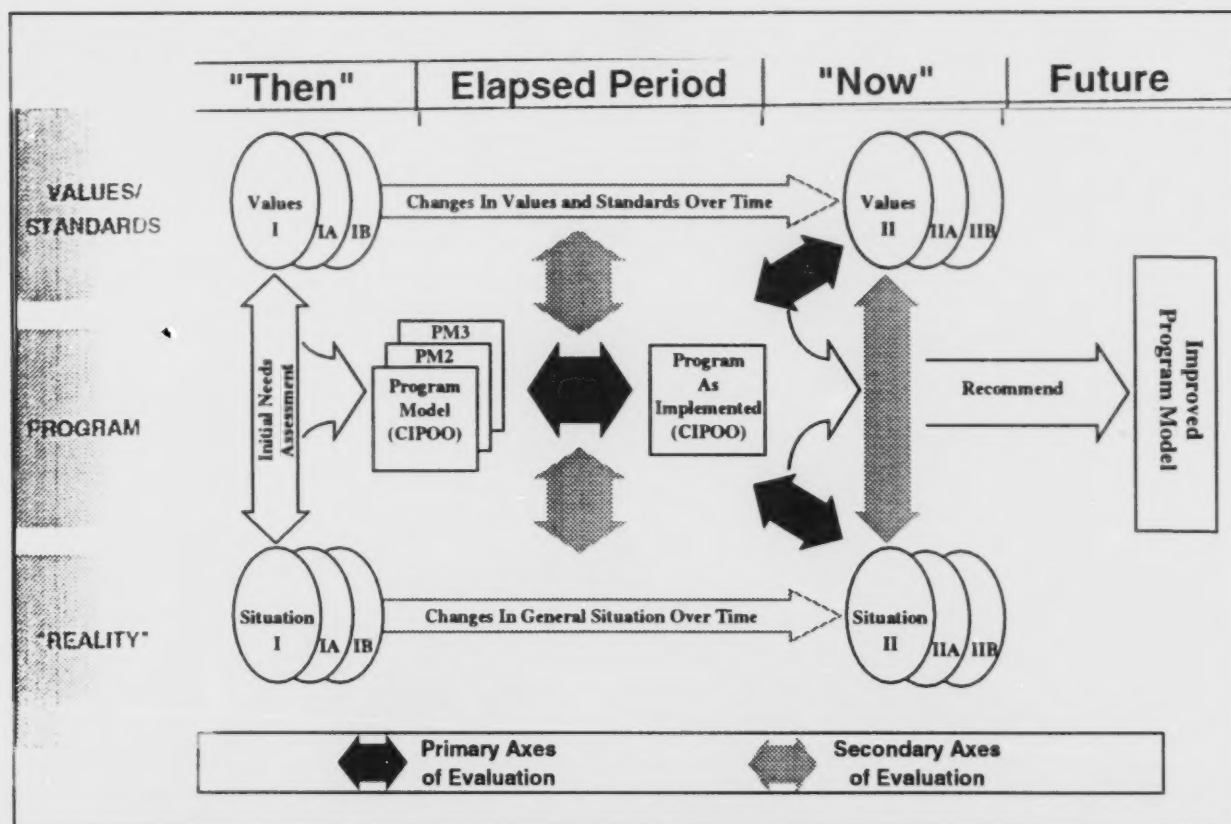


Fig. 4.4 Flow diagram of evaluation comparisons

impact of other influences as well. At the moment when evaluation is undertaken ("now" in Figure 4.4), some combination of program personnel, program beneficiaries and outside consultants examines the degree of implementation of the model, assesses impact by comparing the current situation of beneficiaries to the previous one and considering the role of the program in these changes, and once again compares the current situation to values and standards, which may also have changed in the meantime. This exercise leads to recommendations for improvements in the program model or strategy. As Rondinelli (1993) has pointed out, most programs and projects are essentially "policy experiments" and evaluation is the function that enables us to draw conclusions from them.

A complementary portrayal of the comparisons underlying program evaluation - this one cross-sectional rather than longitudinal in nature and drawn from the work of Robert Stake (cf. Worthen 1973: 106-123) - is given in Figure 4.5. Stake's scheme highlights the way in which evaluators contrast varied dimensions of the "program-as-planned" with data on the "program-as-implemented" in order to derive suggestions for a revised or improved program model - and how they conduct internal analyses and comparisons among context, input, process, output and outcome information in order to identify the strong and weak points of the model and of its implementation record that may explain observed results.

Stake calls internal comparisons of dimensions of the program-as-planned "analyses of logical consequence," because they involve considering to what degree a program could reasonably expect to accomplish the targeted results and impacts with the proposed inputs and processes, and given its own understanding of context. (This is much like the sort of *ex ante* evaluation or feasibility analysis of a proposed program that decision-makers and planners must often conduct before ever choosing to implement it.) He calls internal comparisons among dimensions of the program-as-implemented "analyses of empirical consequence". In other words, once the data are in hand, the evaluator is investigating what results were obtained from the inputs actually invested in the program and the processes actually used, given the real context within which it had to operate.

Both means of organizing and comparing data on program operations may prove useful to the evaluator. Of course, there is no obligation to perform such exercises with actual tables of the sort pictured. The

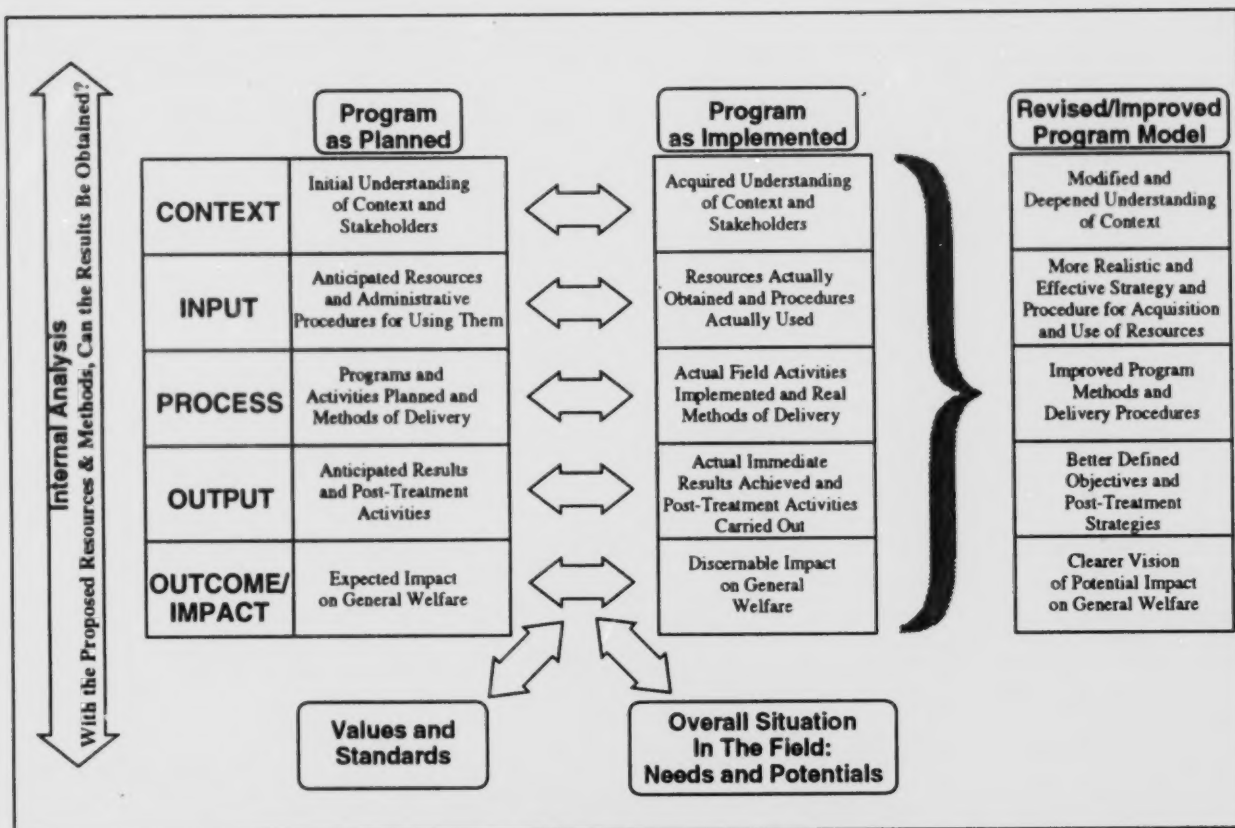


Fig. 4.5 Stufflebeam scheme of evaluation comparisons

table is mostly an image meant to evoke the way in which perceptions and observations concerning resources, methods and results can be aligned in the evaluation of a program. Whatever methodology, metaphor or image best serves evaluation purposes in the particular case at hand and helps to facilitate stakeholder understanding of the issues involved is the appropriate one in those circumstances. But keep in mind that evaluation is as much a mental exercise as it is an accountability mandate or a research endeavor. To work together and participate in a new undertaking, people need a common set of images, even if they may - and perhaps should - be discarded once the exercise is over.

Conclusions: Learning to paint the whole picture

In all the charts and explanations of the preceding sections, the underlying issue is simply what means to use for getting the "whole picture" of the program to be evaluated, for helping stakeholders to get it as well, and for understanding the images of it that they already have. Conduct of an evaluation, and particularly a participatory evaluation, supposes and requires some sort of common - or at least overlapping - language for talking about what was intended, what has been done, and what can be learned from the experience: in short, a shared set of representations of a program.

This is not to say that all stakeholder groups will see the program or its environment in the same way, quite the reverse. And these divergences are important to note, to respect and to analyze. But to be able to capture and compare them, one needs some common frame of reference: a stage, so to speak, on which data and differing stakeholder viewpoints can be presented, examined and discussed. Offering tools for constructing that stage has been the purpose of this chapter.

In any case, "systems thinking" about the dimensions and contexts of nonformal education - that is, an aptitude for picturing the entire program, its sources and consequences, within its environment and against the backdrop of the related habits and institutions that exist already - is more important than any particular model or scheme for representing a program. It is a framework to which we will frequently refer in the rest of this book.

Notes

¹ "Should" is a key word here. In fact, there is a tendency for discussion and assessment to focus on what is easily quantifiable, because it fits most neatly into the boxes. Self (1985) speaks in the US Congressional context of what legislative staffers refer to as the problem of "horse and rabbit stew": a little of the former inevitably colors the taste of the entire dish. They mean by this that numbers have such respect in the culture of government and business - a phenomenon to be considered more closely in the next chapter - that not only are they requested of each deponent, but as soon as they enter the debate participants tend to neglect all the qualitative and contextual detail that is needed or available for decision-makers.

² Note that one can equally well proceed in the reverse direction, starting with current resources and personnel and asking what their functions are, then seeking to articulate what ends the services or materials they provide are supposed to serve and how these ends are accomplished. In this manner one moves "left to right" on the systems diagram and eventually comes to program outputs and outcomes. Actually, it is possible to start at any point in the scheme and "iterate" several times in the attempt to fill in the overall picture. Whatever method proves stimulating for the participants and serves to make thought flow is the appropriate one.

³ Note that the most interesting contrasts in the evaluation of a program which has an explicit program model or theory will probably crop up in the comparison between that theoretical construct or program model on one hand and the "balance sheet" of what has actually transpired in the field on the other. In the case, however, of a program that "just grew" from spontaneous initiatives without any particular plan or advance design, the revelations are just as likely to come from the reconstitution of the implicit program model itself. Since this will be essentially a representation of what is happening in the field, as opposed to a portrait of some prior theory about how to do things, the analysis of program methodology will essentially lie in confronting the incoherencies in these actual procedures and asking how they might be resolved and the entire process made more effective. ...Or, as one university professor is seen remarking to another in a cartoon that recently appeared in the United States, "That's all very well in practice, but will it work out in theory?"

References

- Argyris, C. and Donald Schon. 1978. *Organizational Learning: A Theory of Action Perspective*. Reading, MA: Addison-Wesley Publishing.

- Gareau, E. 1995. "The Educational Implications of Catholic Social Action in Latin America: A Case Study of the APOV Program in Brazil". Unpublished dissertation: Florida State University, Tallahassee, Florida.
- Rondinelli, D.A. 1993. *Development Projects as Policy Experiments: An Adaptive Approach to Development Administration*. 2nd ed. London, New York: Routledge.
- Scriven, M. 1973. "Goal-Free Evaluation" In: E.R. House ed. *School Evaluation: The Politics and Process*. (319-328). Berkeley, CA: McCutchan.
- Stufflebeam, D. et al., 1971. *Educational Evaluation and Decision-Making in Education*. Itasca, IL. Peacock.
- Worthen, B.R. 1973. *Educational Evaluation: Theory and Practice*. Worthington, Ohio: C.A. Jones Publishing.

Part III

**Technical Dimensions
of Evaluation Design**

CHAPTER 5

DATA QUALITY

"Data" are the raw materials from which evaluations are typically built. Arguably they are to the evaluator what wood or brick or stone are to the builder. And they may be of good or bad quality - just as a house of brick or stone is generally stronger and more durable than one made of straw or cardboard. What determines the "quality" of data?

To answer that question, we must ask for a moment what "data" themselves are. And that inquiry will lead us into questions of epistemology (or theories of knowledge) and the philosophy of science.

How do we "know" something, and how is knowledge created? In other words, how do we discover things and by what means do we demonstrate their truth? Epistemology deals with the first issue in each of the foregoing sentences (the nature of knowledge) and the philosophy of science with the second (the demonstration or establishment of truth). There are varying answers to these questions, and the positions that one takes with respect to them - whether consciously or unconsciously - have very important consequences for evaluation design.

The philosophical stance of evaluators

Guba and Lincoln (1994:105) have suggested that the maze of differing schools of thought and points of view regarding these underlying philosophical matters can be helpfully categorized into four groups, two of them representing currently dominant positions and the other two alternative positions or challenges to the mainstream.

The two dominant philosophies are generally referred to as positivism and post-positivism. Positivism is the traditional philosophy of Western science, post-positivism a modification developed to accommodate an increased sense of uncertainty about the "laws" of the universe. Both schools hold that there is a single objective reality out there and that the purpose of human inquiry is to determine its characteristics. Positivists maintain that reality is undergirded by immutable natural laws and objects that systematic scientific inquiry (and only the proper kind of scientific inquiry) can reveal and that are independent of context or

human perception. These natural laws are "transcendent" in the sense that they do not vary with time, place or circumstances.

Post-positivists soften this stance by admitting that reality is only imperfectly apprehended by the flawed procedures of human research and that circumstances may modify our perception of it. Under one set of conditions we may perceive one aspect of reality; under different circumstances another, perhaps overlapping, one. They are consequently more open to the use of multiple methods of inquiry as a means of compensating for weaknesses in human perception and to the importance of including context in research on natural phenomena as a source of understanding.

On the other side of the coin are two alternate points of view that reject to some extent the notion held in common by the dominant schools that objective reality is a given entity to be apprehended in potentially unbiased or complete fashion by human inquiry. Both of these schools - critical theory and constructivism - point out that all observation and inquiry are irremediably colored by the perspective of the observer, since reality can only be perceived from some particular perspective. They do not so much deny the reality of the external world as they question our ability to grasp it in "unbiased" manner.

Critical theorists admit that reality may in theory be directly understandable, but assert that the evolution of human societies and the formation of cultures and of the ideologies that support power have created institutional patterns that condition our perceptions and are taken for "real" even though they are actually pure products of human consciousness and effort. (Remember that, in sociological parlance, the term "institution" refers not just to large formal organizations, but to any stylized pattern of social behavior - like prescribed rites of passage - that a group of people have adopted and that they predictably reproduce in their interactions.) For example, most cultures assign and institutionalize a particular range of roles and attributes for men and women. Whatever a woman may "objectively" be capable of learning and doing, her capacities are largely perceived through the filter of this subtle cultural screen. The purpose of research in the critical theory tradition is therefore essentially to help people understand the historical and contingent nature of these ways of perceiving the world and, when necessary, to challenge and change both the ideologies and the social structures that they support.

Constructivists stress that the multiple facets of reality are apprehended in different ways by different groups of people, who reconstruct these perceptions into a view of the world that works for them. For constructivists, there is no single "reality" out there that science can unequivocally perceive. We build versions and visions of reality out of our own sense perceptions, our own meaning schemes and needs: we in effect "construct" our own knowledge, and to do so is the only way to learn. The challenge of research in the constructivist tradition is to uncover these personal and social constructions of reality and to merge and build them, through consensus, into larger and more inclusive views that offer the basis for a more just society. In essence, constructivist insist that we make our *own* reality, whether consciously or unconsciously. That point of view may in turn radically alter the meaning of notions like "data," "objectivity," and "reliability."

The two alternate schools have scored some telling points in the debate, and though positivism continues to hold sway both in the popular image of science and in most of the major institutions responsible for funding research and knowledge acquisition, it is no longer entirely tenable on philosophical grounds. To a certain extent, post-positivism can be seen as an attempt to salvage what one could from its collapse and not to "throw the baby out with the bath water."

Consequences for evaluation

What differences do these issues make for the practical concerns of evaluators and for our understanding of the issues surrounding "data quality"? There are obviously important disagreements between those who say that "program results" are a distinct and objectively verifiable phenomenon that evaluation is designed to measure, and those who insist that there are no results independent of stakeholders' conceptions of them and that the notion of "results" may be conceived (and so measured) very differently from one such group to another. In designing a program evaluation, a positivist naturally places emphasis on the reliability of data and the quality of standardized instruments, whereas a constructivist may be equally interested by what different groups consider to be relevant "data" in the first place and how each has built its own view of the program. In their co-authored work on evaluation methodology, *Fourth*

Generation Evaluation, Guba and Lincoln (1989) in fact argue that evaluations must be "constructivist" in orientation in the sense that they should be built up directly from stakeholder perceptions and concerns.

The meaning of "data"

What then shall we mean by the term "data"? Everyone knows in a common sense fashion what is being talked about. Data are all those different kinds of information from which evaluations and research studies are built and that must somehow be assembled by evaluators or researchers. But if there is so much fundamental philosophical disagreement between different schools or camps of researchers, is there any common topic to discuss?

The rest of this chapter should help to answer that question and to examine in more practical terms how evaluators go about ensuring that these "building blocks" of their studies are as solid and trustworthy as possible. But it is not too early to try discerning some common ground on which to stand. The various parties to the debate seem at least to agree that in the act of inquiry there are two sides -- one represented by the observer and the other by the situation one is observing or in which one is immersed.¹ Furthermore, for all parties involved, "data" arguably represent some sort of trace or record of this encounter between observer and situation -- and the methodology of research essentially involves cultivating and organizing these encounters, and then portraying their results, in a manner that is demonstrably faithful to the purposes of the inquiry and both understandable and usable for the community most concerned by it. Disagreements arise over the question of how much of the observer must, does or should go into the portrayal and to what degree the observer should and can be part and parcel of the situation or, on the contrary, removed from it.

We will return to these considerations in the chapter on qualitative evaluation technique, for qualitative methodology is designed in part to capture and make use of the more complex and differentiated notion of information that drives the alternate philosophies of evaluation and research. In this chapter, our purpose is to examine what factors determine the quality and utility of the data or information gathered during an evaluation. It is at least evident that the matter will be viewed differently depending on the philosophical position of the evaluator, and of the

*Box 5.1. What differences it makes:
The example of educational sector assessment*

Since its entry into the field of education in the 1970s, the World Bank has been a proponent of the method of educational sector assessment. Sector assessment is a type of macro-evaluation of a country's (or a region's) educational system in which a team of experts assembles data on resources, methods, outputs and impacts of the different types and cycles of education that make up the overall system, compares them with information on the economy's manpower needs and the government's fiscal capacity, and draws conclusions regarding priority areas for investment in educational improvement.

Though much useful data (most of it second-hand) is gathered and analyzed in this process and the assessment is given a highly "objective" aura, critics have pointed out that these sector assessments are typically all framed in a positivistic and econometric manner that tends to support structural readjustment policies of the Bank Group and leaves little place for consideration of the cultural, historical and political dimensions of education or the potentially divergent points of view of different stakeholder groups. This particular perspective may then be virtually "imposed" on the host country through loan conditionalities. A number of sessions at Comparative International Education Society meetings in the United States have been devoted over the last few years to the issue and to proposals that sector assessment teams include as a matter of course both host country nationals and outside "experts" with differing paradigmatic viewpoints. (cf. Easton 1988)

agencies or persons funding evaluation work. (Some of these differences are illustrated in the example in Box 5.1). In the rest of this chapter, therefore, we will attempt to view each issue from two perspectives: that of currently dominant schools, and that more characteristic of alternative paradigms. This is in itself a form of triangulation!

Issues of accuracy

Traditionalists' concerns with data quality generally revolve around two points: (a) gauging and reducing the "margin of error" in the data

collected; and (b) laying a basis for good judgment. Their prescription for alleviating these concerns is essentially to develop correct research procedures.

Appreciating and reducing the margin of error

Even traditionalists readily admit that all information contains elements of error. We nearly never get 100% accurate information about the issues that most interest us or concerning the entire population that we wish to understand. As the poet Robert Frost once put it --

We dance around in a circle and suppose,
But the secret sits in the center and knows.

Traditionally, three elements of this inevitable margin of error are distinguished: reliability, validity and representativity (sampling error). In a sense, all could be considered a version of sampling error, a point to which we shall return.

Reliability

Reliability refers to the difference between "objective reality" - e.g. the true value of some parameter - and the version of it that we are able to apprehend with our instruments or measurement techniques. There is some degree of error in every attempt at measurement; the problem comes in determining what kind and how much. For example, a defective scale may give a reading of someone's weight that is off by some undetermined amount; or it may give different readings at different moments, depending on the air temperature or on whether you stand on the front or the back of the plate.

The effort to assess error is obviously beset with some problems. How can we even define a "true value," since we don't know exactly what this reference point is? Statisticians have adopted a way around the obstacle: this "true" parameter is defined as the average value that we would obtain by attempting an infinite number of measurement efforts (e.g., on an infinite number of different scales) and then calculating their central tendency. It is supposed that our errors in this case would be randomly distributed: those on the "up" side would cancel out those on the "down" side and the average would get progressively closer to the

true value. Reliability is therefore sometimes estimated by determining the degree of fluctuation in the information on a given phenomenon (for example, a student's knowledge of arithmetic) obtained by similar instruments used at different points in time or by different individuals simultaneously.

Validity

Validity, on the other hand, refers to the discrepancy between what we are able to measure or observe and the topics or concepts that really interest us. There are numerous instances where we cannot directly observe the characteristics or subjects of greatest concern to an evaluation or a research study. This is obviously the case with respect to mental aptitudes and attitudes, but equally so in regard to a host of behaviors that take place in locations, at times or in ways that are very difficult to observe. Suppose, for example, that we wished to estimate the extent of use of some medicine as malarial prophylaxis among a given group of people. We might be able to determine quite precisely - i.e., with a high degree of "reliability" - the proportion of the group that *knew* the correct dosage or the proportion that considered malaria both a serious and preventable enough problem to merit prophylactic effort. But neither of these measures is the same as the topic that really concerns us: namely, what proportion of the group actually *uses* the medicine in question as malarial prophylaxis.

A number of the other subjects of interest in NFE evaluations - agricultural knowledge, civic spirit or volunteerism, value placed on girls' education - are equally difficult to observe directly. We can only get at them by defining a series of "indicators" or related phenomena (like questionnaire results or observable behaviors) that are more easily measurable and are closely enough correlated with the level of the attribute of interest so that they provide a plausible barometer of its intensity. As soon as we do so, however, the problem of their validity arises. Since these measures are not identical to the subject of concern, how close are we to what we want to know and what degree of error has the indirect procedure of measurement introduced?

In fact, the different philosophical schools to which we have referred use different versions of the notion of the validity or "plausibility" of concepts. In short, they give different answers to the question "What makes a theory or a construct 'plausible'?" Traditionalists - positivists

and post-positivists - hold that it is the consensus or endorsement of the scientific community (and thus of a trained intellectual elite) that establishes the plausibility of concepts. This consensus is supposedly determined in large measure by the degree to which "correct" inquiry procedures have been observed - "correct" meaning in conformity with accumulated methodological tradition - but it may of course also be influenced by subtle biases or interests held in common by the scientific community itself. Critical theorists and constructivists, on the other hand, lean instead toward a notion of "discursive validity" in which the worth of constructs or truth claims is established or negotiated by stakeholders as a function of their different but complementary viewpoints. From this viewpoint, "plausibility" is an evolving phenomenon, a moving target whose representations are always *ad hoc* and temporary.

Representativity

Representativity or sampling error concerns the discrepancy between the characteristics of those units or people on whom we are able to gather information and the "population" of such entities that actually interests us. Generally evaluations are intended to give a good estimate of the situation of entire programs or population groups. Yet the actual surveys or inquiries can only be carried out with a fraction of this group and during only a fraction of the time over which the program takes place.

Even if our measurements are entirely reliable and valid - that is, even if there is no instability in our estimates and what we are observing is identical to the phenomena of interest to us - considerable error may enter into our results due to the difference between the characteristics of the group that we have surveyed and those of the entire population of interest, and/or between what we are able to observe at those particular times when inquiry took place and what happens at other seasons or periods.² (Box 5.2 gives a classic example of this phenomenon).

You may have noticed that there is an important common denominator to all these definitions or components of the margin of error: In all such cases, we are talking about the discrepancy between, on the one hand, what one might discover with a very large number of trials, with a very great number of different conceptualizations, or by effectively reaching the quasi-totality of the population of interest - and, on the other hand, what we can determine with the limited means at our disposal. In short, in a mechanical sense, each case could be thought of as a sampling

*Box 5.2 Misunderstanding the nature of a sample:
The 1936 Presidential election poll in the United States.*

The classic example of representativity or sampling error in my country comes from the 1936 US presidential election when Alfred Landon, Republican, was running against the incumbent president, Franklin Delano Roosevelt, a Democrat. Roosevelt was resented by many of the well-to-do because he had instituted relatively costly social programs and new controls on industry to soften the effects of the Depression, then at its height. Public opinion polls, conducted in a regular and systematic manner before the election, repeatedly forecast that the challenger Landon would win the election by a comfortable margin. When the election tallies were in, however, the result was just the reverse: Roosevelt had beaten his rival badly.

What had gone wrong with the polls? It turned out that the pollsters had constituted very proper random samples of interviewees...by using the telephone directories as sampling frames. At that time, however, only relatively wealthy people in fact had telephones; few poor or working households did. Consequently, the survey results gave a very accurate picture of the voting intentions of the well-to-do population, but a very poor one of the overall body of voters. The group surveyed was not correctly drawn from, and did not adequately represent, the real population of interest.

problem: out of all possible measures or trials or population units, we have only been able to undertake a restricted number, and we would like to estimate how far the results we obtain with these limited means are likely to wander from the "truth." Analyses and calculations of the margin of error are based on precisely this logic.

This point leads to another: Both in theory and in very practical terms, one should be able to collapse all three elements of the margin of error into one expression or estimation. If we are concerned to estimate the proportion of primary school teachers in a given region of the country who are competent enough in mathematics to handle the accounting of a local cooperative and are trying to do so by means of a test of accounting knowledge administered to a sample of them, all three elements of the margin of error may enter into play: the reliability of scores on the exam, the validity of these data as an indicator of practical accounting competence, and the representativity of the results with respect to the overall population of primary school teachers. One would

have to combine some estimate of all three to get an idea of how far off the proportion derived from test results likely is with respect to the true proportion of competent teachers in the region that interests us.

In practice such efforts are seldom made.³ But at least this exposition helps to make us sensitive to the different sources of error inherent in evaluation design and in the data produced. And a good deal of the artistry of evaluation design is traditionally involved with foreseeing and minimizing potential sources of error in instrument construction, data collection and data interpretation. The example in Box 5.3 summarizes the inter-relatedness of the three sources of error we have been discussing in a practical setting.

Alternate perspective

Most of the whole preceding discussion is based on the notion that there is some unequivocal "reality" out there that we are trying to determine, some truth regarding our evaluation questions or variables that serves as the criterion for judging the reliability and validity of our efforts. Evaluators with a critical theory or constructivist perspective object that this notion has little meaning in and of itself, since data can *never* be separated - never, at least, by some technical procedure - from the multiple perspectives that affect their collection or the elements of ideology interwoven in their definition. As Gephart (1988) has pointed out, uses of quantitative data finally have qualitative foundations: just what numbers mean and the uses to which they are put vary from culture to culture. For the notion of reliability, critical theorists and constructivists might therefore substitute one of usable and maximally-inclusive social consensus. From this point of view, what makes the conclusions of evaluation or research "trustworthy" is the fact that stakeholders (including the researchers) have fully debated them and come to an agreement that can serve as a basis for common action.

Even if our data and evaluation results are essentially social constructions rather than transparent renderings of some unequivocal reality, from a constructivist perspective that in no way diminishes their worth. Social construction is a critically important function. Consensus about what we shall treat as reality is a necessary starting point for collective action. What is essential from this viewpoint is less to measure discrepancies on a scale than to appreciate the difference between the viewpoints of different stakeholders about the truthfulness and utility of

*Box 5.3. Triple sources of error in evaluation data:
An example from training in oral rehydration therapy (ORT)*

Evaluation of the competencies actually acquired by women who have taken part in training in ORT could be subject to the same three inter-related sources of error highlighted in the foregoing pages, particularly if assessment is performed with the aid of some written instrument or questionnaire and beneficiaries of the program are numerous enough - or evaluation resources limited enough - so that the assessment can only be carried out with a sample of women. Suppose that the method used for assessing these competencies is to question selected graduates of the training cycle on the steps in rehydration therapy. Three issues then arise:

1. How *reliably* does the procedure determine the level of knowledge of ORT therapy of the women in the sample? Do they and the evaluators understand the terminology used and is the procedure simple and inviting enough to engage their attention? If not, the results may be due in part to failings in this procedure and would doubtless fluctuate over different administrations: a source of reliability error.
2. How *validly* do these measures of the knowledge of the women indicate the degree to which they can and will apply them in practice? Theoretical knowledge and actual use are not the same thing, though they may be closely related and the first prerequisite to the second. Discrepancies between them are sources of validity error in this example.
3. Finally, how *representative* are the women actually interviewed of the entire group that took the training and/or to which evaluators would like to generalize the results? A sample is never 100% representative of the population from which it is drawn and, depending on the manner in which it is created and the success with which sample members are actually contacted, may be nearer or farther from that goal. And these discrepancies are a source of sampling error.

All three sources of error combine when one wishes to use the results of this particular assessment to characterize the level of practical competence of the entire group of trainees in ORT. The results may still be highly useful, but the possible discrepancies highlighted above and some sense of the related margin of error need to be kept in mind when using them.

different sets of data and to seek a common ground that will serve the purposes of program improvement.

Accounting for the level of error

Whatever the quality of evaluation data, their *purpose* is to provide a basis for decision making. Data that contain elements of error - as all data do - can still be helpful for program improvement ends as long as one is clear about their quality and the uses to which they can and cannot be put. The fundamental issue here therefore concerns how data are actually used in reaching decisions and conclusions, in formulating judgments about programs and their most salient needs.

One traditional approach to these issues is to estimate the margin of error in data and to take account of the level of precision needed for decision purposes. Two conventions from the physical sciences are of material assistance at this point: the notion of "significant figures" and that of "orders of magnitude."

The notion of significant figures

The degree of precision of numerical data is generally indicated in scientific notation by what is called the number of significant figures. To write that there are 32,432 people in a given city would mean, from this perspective, that we are certain there are exactly that many - neither 32,431 nor 32,433. If in fact our estimates are only accurate within five units or so, then the last figure in the population estimate should be rounded off to zero.⁴ Writing "about 32,430" would, in this sort of notation, signify that we are at least sure the real value falls somewhere between 32,425 and 32,434. Similarly if our estimates were only accurate to the nearest 100, we would write "about 32,400." If they were accurate only to the nearest thousand, we might write "about 32,000." To write "about 30,000" would indicate an even grosser level of estimation, accurate only to the nearest 10,000. Any greater inaccuracy would make it no longer worthwhile to offer any quantitative estimate whatsoever.

These distinctions are important, because in fact we are seldom sure of statistics to the very last digit, and we need a handy fashion to remind ourselves and others of this and of our level of certitude. One of the places where this counsel is the most regularly ignored is in the calculation and reporting of percentages. If 42 out of 67 nurse practitioners in

*Box 5.1. What differences it makes:
The example of educational sector assessment*

Since its entry into the field of education in the 1970s, the World Bank has been a proponent of the method of educational sector assessment. Sector assessment is a type of macro-evaluation of a country's (or a region's) educational system in which a team of experts assembles data on resources, methods, outputs and impacts of the different types and cycles of education that make up the overall system, compares them with information on the economy's manpower needs and the government's fiscal capacity, and draws conclusions regarding priority areas for investment in educational improvement.

Though much useful data (most of it second-hand) is gathered and analyzed in this process and the assessment is given a highly "objective" aura, critics have pointed out that these sector assessments are typically all framed in a positivistic and econometric manner that tends to support structural readjustment policies of the Bank Group and leaves little place for consideration of the cultural, historical and political dimensions of education or the potentially divergent points of view of different stakeholder groups. This particular perspective may then be virtually "imposed" on the host country through loan conditionalities. A number of sessions at Comparative International Education Society meetings in the United States have been devoted over the last few years to the issue and to proposals that sector assessment teams include as a matter of course both host country nationals and outside "experts" with differing paradigmatic viewpoints. (cf. Easton 1988)

agencies or persons funding evaluation work. (Some of these differences are illustrated in the example in Box 5.1). In the rest of this chapter, therefore, we will attempt to view each issue from two perspectives: that of currently dominant schools, and that more characteristic of alternative paradigms. This is in itself a form of triangulation!

Issues of accuracy

Traditionalists' concerns with data quality generally revolve around two points: (a) gauging and reducing the "margin of error" in the data

collected; and (b) laying a basis for good judgment. Their prescription for alleviating these concerns is essentially to develop correct research procedures.

Appreciating and reducing the margin of error

Even traditionalists readily admit that all information contains elements of error. We nearly never get 100% accurate information about the issues that most interest us or concerning the entire population that we wish to understand. As the poet Robert Frost once put it --

We dance around in a circle and suppose,
But the secret sits in the center and knows.

Traditionally, three elements of this inevitable margin of error are distinguished: reliability, validity and representativity (sampling error). In a sense, all could be considered a version of sampling error, a point to which we shall return.

Reliability

Reliability refers to the difference between "objective reality" - e.g. the true value of some parameter - and the version of it that we are able to apprehend with our instruments or measurement techniques. There is some degree of error in every attempt at measurement; the problem comes in determining what kind and how much. For example, a defective scale may give a reading of someone's weight that is off by some undetermined amount; or it may give different readings at different moments, depending on the air temperature or on whether you stand on the front or the back of the plate.

The effort to assess error is obviously beset with some problems. How can we even define a "true value," since we don't know exactly what this reference point is? Statisticians have adopted a way around the obstacle: this "true" parameter is defined as the average value that we would obtain by attempting an infinite number of measurement efforts (e.g., on an infinite number of different scales) and then calculating their central tendency. It is supposed that our errors in this case would be randomly distributed: those on the "up" side would cancel out those on the "down" side and the average would get progressively closer to the

true value. Reliability is therefore sometimes estimated by determining the degree of fluctuation in the information on a given phenomenon (for example, a student's knowledge of arithmetic) obtained by similar instruments used at different points in time or by different individuals simultaneously.

Validity

Validity, on the other hand, refers to the discrepancy between what we are able to measure or observe and the topics or concepts that really interest us. There are numerous instances where we cannot directly observe the characteristics or subjects of greatest concern to an evaluation or a research study. This is obviously the case with respect to mental aptitudes and attitudes, but equally so in regard to a host of behaviors that take place in locations, at times or in ways that are very difficult to observe. Suppose, for example, that we wished to estimate the extent of use of some medicine as malarial prophylaxis among a given group of people. We might be able to determine quite precisely - i.e., with a high degree of "reliability" - the proportion of the group that *knew* the correct dosage or the proportion that considered malaria both a serious and preventable enough problem to merit prophylactic effort. But neither of these measures is the same as the topic that really concerns us: namely, what proportion of the group actually *uses* the medicine in question as malarial prophylaxis.

A number of the other subjects of interest in NFE evaluations - agricultural knowledge, civic spirit or volunteerism, value placed on girls' education - are equally difficult to observe directly. We can only get at them by defining a series of "indicators" or related phenomena (like questionnaire results or observable behaviors) that are more easily measurable and are closely enough correlated with the level of the attribute of interest so that they provide a plausible barometer of its intensity. As soon as we do so, however, the problem of their validity arises. Since these measures are not identical to the subject of concern, how close are we to what we want to know and what degree of error has the indirect procedure of measurement introduced?

In fact, the different philosophical schools to which we have referred use different versions of the notion of the validity or "plausibility" of concepts. In short, they give different answers to the question "What makes a theory or a construct 'plausible'?" Traditionalists - positivists

and post-positivists - hold that it is the consensus or endorsement of the scientific community (and thus of a trained intellectual elite) that establishes the plausibility of concepts. This consensus is supposedly determined in large measure by the degree to which "correct" inquiry procedures have been observed - "correct" meaning in conformity with accumulated methodological tradition - but it may of course also be influenced by subtle biases or interests held in common by the scientific community itself. Critical theorists and constructivists, on the other hand, lean instead toward a notion of "discursive validity" in which the worth of constructs or truth claims is established or negotiated by stakeholders as a function of their different but complementary viewpoints. From this viewpoint, "plausibility" is an evolving phenomenon, a moving target whose representations are always *ad hoc* and temporary.

Representativity

Representativity or sampling error concerns the discrepancy between the characteristics of those units or people on whom we are able to gather information and the "population" of such entities that actually interests us. Generally evaluations are intended to give a good estimate of the situation of entire programs or population groups. Yet the actual surveys or inquiries can only be carried out with a fraction of this group and during only a fraction of the time over which the program takes place.

Even if our measurements are entirely reliable and valid - that is, even if there is no instability in our estimates and what we are observing is identical to the phenomena of interest to us - considerable error may enter into our results due to the difference between the characteristics of the group that we have surveyed and those of the entire population of interest, and/or between what we are able to observe at those particular times when inquiry took place and what happens at other seasons or periods.² (Box 5.2 gives a classic example of this phenomenon).

You may have noticed that there is an important common denominator to all these definitions or components of the margin of error: In all such cases, we are talking about the discrepancy between, on the one hand, what one might discover with a very large number of trials, with a very great number of different conceptualizations, or by effectively reaching the quasi-totality of the population of interest - and, on the other hand, what we can determine with the limited means at our disposal. In short, in a mechanical sense, each case could be thought of as a sampling

*Box 5.2 Misunderstanding the nature of a sample:
The 1936 Presidential election poll in the United States.*

The classic example of representativity or sampling error in my country comes from the 1936 US presidential election when Alfred Landon, Republican, was running against the incumbent president, Franklin Delano Roosevelt, a Democrat. Roosevelt was resented by many of the well-to-do because he had instituted relatively costly social programs and new controls on industry to soften the effects of the Depression, then at its height. Public opinion polls, conducted in a regular and systematic manner before the election, repeatedly forecast that the challenger Landon would win the election by a comfortable margin. When the election tallies were in, however, the result was just the reverse: Roosevelt had beaten his rival badly.

What had gone wrong with the polls? It turned out that the pollsters had constituted very proper random samples of interviewees...by using the telephone directories as sampling frames. At that time, however, only relatively wealthy people in fact had telephones; few poor or working households did. Consequently, the survey results gave a very accurate picture of the voting intentions of the well-to-do population, but a very poor one of the overall body of voters. The group surveyed was not correctly drawn from, and did not adequately represent, the real population of interest.

problem: out of all possible measures or trials or population units, we have only been able to undertake a restricted number, and we would like to estimate how far the results we obtain with these limited means are likely to wander from the "truth." Analyses and calculations of the margin of error are based on precisely this logic.

This point leads to another: Both in theory and in very practical terms, one should be able to collapse all three elements of the margin of error into one expression or estimation. If we are concerned to estimate the proportion of primary school teachers in a given region of the country who are competent enough in mathematics to handle the accounting of a local cooperative and are trying to do so by means of a test of accounting knowledge administered to a sample of them, all three elements of the margin of error may enter into play: the reliability of scores on the exam, the validity of these data as an indicator of practical accounting competence, and the representativity of the results with respect to the overall population of primary school teachers. One would

have to combine some estimate of all three to get an idea of how far off the proportion derived from test results likely is with respect to the true proportion of competent teachers in the region that interests us.

In practice such efforts are seldom made.³ But at least this exposition helps to make us sensitive to the different sources of error inherent in evaluation design and in the data produced. And a good deal of the artistry of evaluation design is traditionally involved with foreseeing and minimizing potential sources of error in instrument construction, data collection and data interpretation. The example in Box 5.3 summarizes the inter-relatedness of the three sources of error we have been discussing in a practical setting.

Alternate perspective

Most of the whole preceding discussion is based on the notion that there is some unequivocal "reality" out there that we are trying to determine, some truth regarding our evaluation questions or variables that serves as the criterion for judging the reliability and validity of our efforts. Evaluators with a critical theory or constructivist perspective object that this notion has little meaning in and of itself, since data can *never* be separated - never, at least, by some technical procedure - from the multiple perspectives that affect their collection or the elements of ideology interwoven in their definition. As Gephart (1988) has pointed out, uses of quantitative data finally have qualitative foundations: just what numbers mean and the uses to which they are put vary from culture to culture. For the notion of reliability, critical theorists and constructivists might therefore substitute one of usable and maximally-inclusive social consensus. From this point of view, what makes the conclusions of evaluation or research "trustworthy" is the fact that stakeholders (including the researchers) have fully debated them and come to an agreement that can serve as a basis for common action.

Even if our data and evaluation results are essentially social constructions rather than transparent renderings of some unequivocal reality, from a constructivist perspective that in no way diminishes their worth. Social construction is a critically important function. Consensus about what we shall treat as reality is a necessary starting point for collective action. What is essential from this viewpoint is less to measure discrepancies on a scale than to appreciate the difference between the viewpoints of different stakeholders about the truthfulness and utility of

*Box 5.3. Triple sources of error in evaluation data:
An example from training in oral rehydration therapy (ORT)*

Evaluation of the competencies actually acquired by women who have taken part in training in ORT could be subject to the same three inter-related sources of error highlighted in the foregoing pages, particularly if assessment is performed with the aid of some written instrument or questionnaire and beneficiaries of the program are numerous enough - or evaluation resources limited enough - so that the assessment can only be carried out with a sample of women. Suppose that the method used for assessing these competencies is to question selected graduates of the training cycle on the steps in rehydration therapy. Three issues then arise:

1. How *reliably* does the procedure determine the level of knowledge of ORT therapy of the women in the sample? Do they and the evaluators understand the terminology used and is the procedure simple and inviting enough to engage their attention? If not, the results may be due in part to failings in this procedure and would doubtless fluctuate over different administrations: a source of reliability error.
2. How *validly* do these measures of the knowledge of the women indicate the degree to which they can and will apply them in practice? Theoretical knowledge and actual use are not the same thing, though they may be closely related and the first prerequisite to the second. Discrepancies between them are sources of validity error in this example.
3. Finally, how *representative* are the women actually interviewed of the entire group that took the training and/or to which evaluators would like to generalize the results? A sample is never 100% representative of the population from which it is drawn and, depending on the manner in which it is created and the success with which sample members are actually contacted, may be nearer or farther from that goal. And these discrepancies are a source of sampling error.

All three sources of error combine when one wishes to use the results of this particular assessment to characterize the level of practical competence of the entire group of trainees in ORT. The results may still be highly useful, but the possible discrepancies highlighted above and some sense of the related margin of error need to be kept in mind when using them.

different sets of data and to seek a common ground that will serve the purposes of program improvement.

Accounting for the level of error

Whatever the quality of evaluation data, their *purpose* is to provide a basis for decision making. Data that contain elements of error - as all data do - can still be helpful for program improvement ends as long as one is clear about their quality and the uses to which they can and cannot be put. The fundamental issue here therefore concerns how data are actually used in reaching decisions and conclusions, in formulating judgments about programs and their most salient needs.

One traditional approach to these issues is to estimate the margin of error in data and to take account of the level of precision needed for decision purposes. Two conventions from the physical sciences are of material assistance at this point: the notion of "significant figures" and that of "orders of magnitude."

The notion of significant figures

The degree of precision of numerical data is generally indicated in scientific notation by what is called the number of significant figures. To write that there are 32,432 people in a given city would mean, from this perspective, that we are certain there are exactly that many - neither 32,431 nor 32,433. If in fact our estimates are only accurate within five units or so, then the last figure in the population estimate should be rounded off to zero.⁴ Writing "about 32,430" would, in this sort of notation, signify that we are at least sure the real value falls somewhere between 32,425 and 32,434. Similarly if our estimates were only accurate to the nearest 100, we would write "about 32,400." If they were accurate only to the nearest thousand, we might write "about 32,000." To write "about 30,000" would indicate an even grosser level of estimation, accurate only to the nearest 10,000. Any greater inaccuracy would make it no longer worthwhile to offer any quantitative estimate whatsoever.

These distinctions are important, because in fact we are seldom sure of statistics to the very last digit, and we need a handy fashion to remind ourselves and others of this and of our level of certitude. One of the places where this counsel is the most regularly ignored is in the calculation and reporting of percentages. If 42 out of 67 nurse practitioners in

our sample respond correctly to a question, it would be a bit abusive and misleading to say that "62.68%" or even "62.9%" responded correctly, even though both of these results can be obtained by division and rounding of the numbers in question. Writing "62.9%" (for example) only makes sense if there is a meaningful difference between 62.8% and 63.0% - which could only happen if there were at least 500 people in our sample, and if we had the requisite level of certainty about these results. In the case just illustrated, it would be more accurate to say that "about 63%" of the nursing personnel in question responded correctly.

The notion of orders of magnitude

To say that numerical estimates are often only approximate is not to say that the less-than-exact ones are useless. To the contrary, we often need no more than approximations for decision purposes, and any attempt to get more precision would be waste effort when an "order of magnitude" or "ballpark" estimate will do. When those responsible for buying supplies for a restaurant learn a week before hand that there will be a special banquet on Saturday night and inquire about how many people will attend, they may only be interested at first in knowing whether one is talking about less than 10 people (a small dinner party), between 10 and 50 (a medium-sized banquet) or over 50 (a really large event). That is enough to let them know how to adjust their buying schedule; the detail can be filled in later. Expending effort at this point arguing over exactly how many will come or calling around town to make a more precise prediction might be a waste of energy that could be better used in other kinds of preparation.

There are many circumstances like this where rough estimates are not only enough, but really offer decision-makers much more than they previously knew. If an agency head wants to make educational television broadcasts to 1500 training centers around the country and has little or no idea how many of them receive signals reasonably clearly, then simply knowing whether one is talking about a handful, several hundred or the vast majority may already be a really valuable piece of information. These are situations in which estimations of the "order of magnitude" of phenomena (e.g., are they in the tens, the hundreds or the thousands?) can prove highly useful. The underlying purpose of studies and assessments, planners are in the habit of saying, is not to capture every dimension of reality but rather to "reduce uncertainty" about the

issues that most concern us.⁵ In any and all cases, it is quite important for evaluators to determine ahead of time just how accurate results need to be to serve the purposes at hand, because increasing accuracy - or every incremental degree in uncertainty reduction - is bought at a price, and therefore at the sacrifice of other things that cannot be done with the same time and resources.

Taken together, these two conventions - the notation of significant figures and the estimation of orders of magnitude - may be of considerable help in making sure evaluators and their audiences respect and remember the margin of error in data. As a rule of thumb, one can either "drop back an order of magnitude" (e.g. from "284" to "about 280") or indicate a margin of error ("280 plus or minus 5 units"). Thus in opinion surveys one will often hear that 43% of the electorate favors such a position, 52% the reverse, with 5% undecided... and, it is added, the results can be taken as accurate within a margin of 3% up or down.⁶

Data quality lies therefore as much in thoughtful estimations of margin of error as it does in accuracy *per se*. Figures that include no information about levels of accuracy are like appliances or tools without user's manuals: one can do a lot of damage with them before getting things right. Qualitative studies accomplish the same end through the technique of "triangulation" - that is, by obtaining descriptions, judgments and assessments of critical phenomena from several different points of view, from several different observers, or via two or more different methods of observation. The discrepancies between these separate accounts not only give the reader of the report a sense of the variability of the data but also furnish information on divergences in the perspectives of the different stakeholder groups involved. Including margins of error in data reports also serves a pedagogical purpose: It can help to sensitize everyone to the indeterminacy of many quantitative results and the consequent importance of supplementing them with qualitative insight.

Alternative perspective

Evaluators of a critical or constructivist frame of mind tend to point out that simply specifying margins of error - quantitative and/or qualitative - does not necessarily lay a basis for good judgment and decision-making. Usually the issue at stake in programmatic decisions is implicitly more

one of *causal attribution* than it is one of simple description. Decision-makers are interested in knowing if achievement levels in one program are significantly better than those in another *because* they want to make judgments about the relative effectiveness of the methods used in the programs in question, or about the relative effectiveness of their staff. Such assessments involve more than accurate description of achievement levels: they involve judgments about the relationship between this information, on the one hand, and data on program methodology or staff performance on the other. In short, the users of evaluation information are more interested in *patterns* of data than in single sets.

The traditional response to this objection is that data should be used to verify or refute *hypotheses* about the relationships between variables and about the causal influences at work - hypotheses quantitatively operationalized in such a manner that the statistical significance of results can be determined. Here too there are strong debates among those of differing philosophical persuasions about the real "significance" of statistical significance and the likelihood of drawing conclusions not biased by the way in which the data have been framed. Critics point out that traditional statistical significance figures only take account of sampling error, and not of the reliability and validity issues that are - as we have seen above - important components of the margin of error.

Post-positivists take a modified tack in their use of hypothesis testing, emphasizing effect size, issues of replicability or accumulation of confirmatory results over time, as well as probability considerations. They have also developed some technical methods for dealing with errors in judgment and for lessening their likelihood and impact. Prominent among these is "sensitivity analysis," which involves running statistical analyses under alternate sets of assumptions about basic parameters in order to see how "sensitive" or "robust" conclusions are to changes (e.g., margins of error) in critical variables. One would want to stick only with a conclusion that was "robust to" the extreme values of the confidence intervals on those variables - i.e., that holds even when the values of the variables are pushed to the outer limits defined by their respective margins of error. This technique essentially amounts to trying out different scenarios or running a series of simulations to see how well the model or the hypothesis stands up to changing assumptions about true values or likely interactions.

The concept of workable truth

These are useful insights but arguably do not address the core issue raised by the alternate perspective. Critical theory proponents and constructivists tend to regard statistical investigation less as an instrument for discovering truth about evaluative questions and causal attributions than as *a language of exposition for developing and presenting points of view*. They maintain that error in judgment is more often due to invalid causal attribution than it is to misapprehension of descriptive magnitudes or error margins - i.e. it is a matter of poor pattern recognition. Data do not have a great deal of substantive significance in and of themselves, only patterns and associations of data do. Data are usually regarded as important (or substantively significant) because they imply a certain association or causal connection among variables of interest.

In the area of pattern recognition, moreover, it is better to be approximately right than exactly wrong. The important thing is to get a workable sense of the "gestalt," of how the system as a whole operates; missing this overall pattern while very accurately locating or describing different pieces of it would serve little purpose. Moreover, patterns - or at least perceptions of them - typically change depending on one's point of view: the moon in the night sky assumes different shapes depending on the position of the earth, grains of sand under a microscope appear radically different from those seen with the naked eye in a desert landscape, and each witness to a crime has a different story. For such reasons, critical theorists and constructivists tend, once again, to place emphasis on the *dialogic process* incorporated in an evaluation - that is, on the way in which contrasting pattern interpretations are solicited and taken into account in analyzing data, presenting them and then drawing conclusions.

In fact, pattern perspectives intervene from the outset. Every datum includes a perspective, if only one asserting that this particular item of information is worth collecting and so worth discriminating from the flow of everyday experience. As courtroom battles over the admissibility of evidence amply demonstrate, choosing and defining the data are themselves highly significant activities. It isn't enough to say, "The facts, ma'am - just the facts." Facts are most often chosen and framed with some perspective in mind - the prosecuting attorney's or the defense's.

We will go more deeply into the issue of causal reasoning in Chapter Twelve on "Impact Evaluation." For the time being, let's come back to the question of *data quality* on which this chapter has focused. We began the chapter by asking what it means to ensure good quality in these basic building blocks of evaluation. Two perspectives were introduced: minimizing margins of error and improving the "usability" of data for decision making. It became evident that data do not speak by themselves, but are used to form patterns and are often defined and derived as part of some preconceived pattern.

This realization does not invalidate all concern with margins of error. Whatever perspective presides over the definition and collection of data, they then acquire a life of their own, pass into circulation and may be put to many different uses. They should always include a "user's manual" indicating how they were derived and what associated margins of error are thought to be. But quality finally inheres in the patterns and judgments built from them more than it does in individual pieces of data. And since multiple patterns can be built from similar sets of data, the issue for improving evaluations concerns particularly how to solicit, compare and confront different perspectives on pattern and causality. The challenge of making evaluation truly "participatory" raised in Chapter Three - that is, of bringing the multiple perspectives of stakeholders to bear on the task - is therefore just as germane to issues of data quality as it is to those of evaluation administration.

A few practical consequences of the points discussed above are worth keeping in mind:

- There are three potential sources of error in data to remember: reliability, validity and sampling error. Awareness of them at least reminds us to make a clear distinction between the results of our inquiries and the "truth," however that is defined. Most statistical significance figures and confidence interval calculations only take account, however, of sampling error.
- From a post-positivist perspective, truth is never entirely knowable. From a critical theory or constructivist perspective, the abstract truth is not worth knowing. Truth is something that we construct and it is imbued with the perspective of our own class and group. Making usable truth means creating dialogue between those with different perspectives.

- Triangulation in research design is the qualitative equivalent of controls on margin of error. It can be usefully paired with a dialogic process in evaluation choreography.
- The whole area of data quality and analysis is rampant with what sociologists call "reification," or the tendency to assume that concepts and norms created by people themselves in their attempt to understand and organize their world are in fact God-given and immutable characteristics of the universe. Ideas are taken for reality as soon as numbers are bestowed on them, and it requires a good bit of

*Box 5.4: Imaginary statistics in Haiti:
Demand creating its own supply*

During an educational sector assessment conducted in the Republic of Haiti in the mid-1980s, the economic analysts on the assessment team were puzzled by a report on manpower demand in the modern sector of the economy (i.e., a quantitative analysis of the likely future needs of industry and government for trained personnel) that was quoted in every planning publication they encountered but, on examination, didn't make much sense. They finally decided to invite the author, who resided in Port-au-Prince, to lunch in order to straighten the matter out.

When they queried him about the apparent inconsistencies in the data of the report, the manpower analyst burst out laughing. "Look," he said, "I tried for weeks to get some reasonable estimate of future skilled labor needs from business leaders and government officials, with absolutely no success. Finally, I just decided to write down something straight out of my head that would be outrageous enough to make them react and provide me with some more valid figures. But no one offered the slightest correction, and since then everyone has been citing the document as if it were gospel truth!"

The story reveals not just that analysts sometimes "cook" their data for the strangest reasons, but also that major institutions of planning and foreign assistance have such a need for numerical justification of their policies - thanks to the great and sometimes blind homage paid to quantitative data as a basis for decision-making - that they may accept anything in reasonable mathematical garb that meets these needs. As has been sometimes said about the use of rate of return figures in educational funding decisions in the United States government, they tend to be sacrosanct precisely because no one understands them!

resolution and courage to put their shaky foundations back into question. Box 5.4 provides a humorous but true example of how strong an influence numbers exert on decision-making.

- One is best advised therefore to be moderately skeptical about definitive solutions for data quality problems, for error and bias are part of all our efforts to measure and describe - and to be a bit suspicious of the failure to evoke these problems. Perhaps the wisest counsel is to treat statistics as a language for debating program results, rather than the ultimate gauge of the worth.

Conclusions

We started this chapter with the notion that good quality data are the fundamental building blocks of evaluation. We have discovered that those building blocks themselves may be quite malleable and subject to varying uses and interpretations. As the 19th century American humorist Mark Twain put it, "Figures don't lie, but liars sure can figure." A positivist orientation to the universe assumes that in evaluation we are simply assembling the hard, irrefutable (and preferably quantified) facts. A critical or constructivist one assumes that we inevitably shape our own perceptions - but can modify those shapes - and it treats data as essentially a language of exposition that we use in developing our own view of a program. These are not necessarily irreconcilable differences; they may be two faces of reality with a degree of natural complementarity. Evaluation cannot proceed, we have seen, without a combination, even a "triangulation" of quantitative and qualitative data.

What is finally important is not so much data as *information* - that is, processed, shared and assimilated impressions of experience that serve to *form* our *internal* view of a program and of the dimensions of social reality that cluster around it, that make that perspective more capacious, comprehensive and functional. Such information usually comes from collective processing of data, from comparing different interpretations and "readings" of this material. For this reason, participatory evaluators typically treat survey or questionnaire results not as necessarily accurate pictures of a program but as "input" to interpretive discussions with stakeholders. Only when these numbers and proportions are enriched by insight do they truly inform.

Quantitative data require positing regularities in program operations, defining, measuring and aggregating them. That is a natural and necessary function of program administration as well as a particular kind of research design, because a program is at root a sort of "regularity" imposed on what might otherwise be disparate, non-comparable or quite simply non-existent processes. Qualitative data, on the other hand, entail attention to the unique properties of each setting and event but provide at the same time a means for discovering underlying similarities among settings, as well as the deeper springs of human action within the setting. In a sense they are no less concerned with regularities, but place emphasis on discovering intrinsic patterns by open admission and appreciation of diversities of form and appearance rather than on imposing those dictated by the program.⁷

Both kinds of data can be aggregated in their own way and used to yield valuable information. Aggregation is in part a political function, a way of "giving voice" to locally-experienced reality in overall characterizations of a program and in decisions about it. Improving data quality, or the usability of available information, therefore fundamentally means improving the accuracy, fluidity and justice with which local perspectives on a program - on its resources, operations, results and ultimate impact - are generated and represented in critical program decisions. Better data quality should be nearly synonymous with better participation.

Notes

¹ There is an extreme philosophical position in the alternate camp -- allied, one might say, with the schools of solipsism and idealism, descended from the work of Bishop Berkeley -- that would maintain that there is no external reality independent of our senses. To the famous question, "If a tree falls in a forest where no sentient being is present, does it make a sound?" Berkeley answered "no," insisting that reality is only an invention or emanation of consciousness. But few evaluators would go so far, however "alternate" their perspective, since evaluation is premised (and funded!) on the assumption that a program independent of the evaluator's perceptions does in effect exist.

² This way of framing the issue of representativity makes it clear that it poses a problem of "temporal sampling" as much as one of what might be called

"demographic" or "geographic" sampling - a question to which we will return in Chapter Seven.

³ Though the exercise is rarely performed, it is possible to combine these three components of the margin of error into an overall estimate. Interestingly, to do so, one should not simply add them up arithmetically, but rather treat them as the sides of a right triangle and their sum as the hypotenuse: that is, square them and then take the square root of their sum.

Sum of reliability and validity error = SRV

$$SRV = \sqrt{(\text{reliability error})^2 + (\text{validity error})^2}$$

$$\text{Overall error} = \sqrt{(SRV)^2 + (\text{representativity error})^2}$$

⁴ By mathematical convention, figures from 1 to 4 are rounded back to the next lower unit, while those from 6 to 9 are rounded up to the next higher unit. As for 5, it is rounded up when the figure immediately to the left (representing the next order of magnitude) is *odd* and is rounded down when that figure is *even*. Thus 276 through 279 become 280 and 271 through 274 become 270. 275 would round up to 280, since the figure immediately to the left (the "7") is odd, whereas 265, for example, would round down to 260, since in that case the figure immediately to the left (the "6") is even.

⁵ From the mathematical point of view, notions of uncertainty and desired level of accuracy in evaluation are based on the perspectives and theories of Bayesian statistics. See, e.g., Pollard 1986.

⁶ In cases like these and in statistical notation in general, the margin of error is sometimes referred to as a "confidence interval" and - as we shall see in the chapter on sampling methods - a level of confidence may even be specified. For example, a statistician may say that there is 95% confidence that the true percentage of those favoring the position in question (estimated at 43%) actually lies between 40% and 46%, and a similar level of confidence that the true percentage of those opposing it (estimated at 52%) lies between 49% and 55%. When these two confidence intervals *overlap* - that is, when the upper bound on one is higher than the lower bound on the other - then we no longer have 95% confidence that the figure that we have estimated as higher is actually above the one that we have estimated as lower. In fact, given the number of undecided voters in our example survey (5%), the results would still be indeterminate even though the confidence intervals do not overlap, because if the real figure for those favoring the referendum were at the upper bound of our confidence interval (i.e. 46%) and all those estimated as undecided ended up voting for it, then the referendum would pass with 51%.

⁷ One should be careful about over-dichotomizing the contrast between qualitative and quantitative data or linking each incontrovertibly to one or another of the philosophical schools mentioned in this chapter. In fact, everyone uses both kinds of information in evaluating programs, and the philosophical disagreements are obviously more complex than simple preferences for a given kind of data. Good qualitative evaluators, moreover, make ample use of numbers; and good quantitative ones understand full well that their figures are necessarily framed by presuppositions and meanings. These issues will be further explored in Chapters Eight and Nine on qualitative evaluation; but for our present purposes some rough distinction between qualitative and quantitative kinds of information do serve to highlight different approaches to the issue of improving data quality in evaluations.

References

- Easton, P.A. 1988. *The Role of Economic Analysis in Educational Policymaking: Case Study of an Educational Sector Assessment in the Republic of Haiti*. Two volumes. Unpublished Ph.D. dissertation. Tallahassee, FL: Florida State University.
- Gephart, R.P. 1988. *Ethnostatistics: Qualitative Foundations for Quantitative Research*. Newbury Park, CA: Sage Publications.
- Guba, E.G. and Y.S. Lincoln. 1994. "Competing Paradigms in Qualitative Research." In: N.K. Denzin and Y.S. Lincoln eds., *Handbook of Qualitative Research*, (105-117). Thousand Oaks, CA: Sage Publications.
- Guba, E.G. and Y.S. Lincoln. 1989. *Fourth Generation Evaluation*. Newbury Park: Sage Publications.
- Pollard, W.E. 1986. *Bayesian Statistics for Evaluation Research: An Introduction*. Beverly Hills, CA: Sage Publications.

CHAPTER 6

INDICATORS AND INSIGHT

The purpose of evaluations, we have said, is to provide insight into the context, operations and effects of social programs. The potential topics are extremely numerous. The notion of "program theory" discussed in Chapter Four suggests that there are multiple dimensions in any education or training program. Though over the long term an evaluation program should certainly develop information and insight concerning the whole spectrum of agency operations and field interventions, most individual evaluation studies concentrate attention on some subset of topics in the interest of economizing resources and focusing efforts on the issues that are of greatest current relevance. These themes and questions are generally encapsulated in the "terms of reference" or objectives of the evaluation and determine much about its design.

Even well-designed themes and questions, however, do not indicate just what data should be collected in the field in order to provide answers and create insight into program operations. A critical step in the design of any evaluation therefore lies in "operationalizing" the themes or topics that have been selected - that is, translating them into specifications about what data should be collected, how it will be gathered, and how it will be analyzed. *When and how* this work of specification is accomplished tends to differ in some important ways between quantitative and qualitative evaluation strategies. In quantitative studies, most of the work is done up front; in qualitative ones, much of it is accomplished in the field in order to allow evaluators to shape the design in response to what they discover as the study develops. Evaluations are often characterized by mixed design: part quantitative, part qualitative. The exact timing of inquiry specification may therefore vary from one component to another of the particular evaluation in question. Certain parts of an evaluation, like achievement tests or closed-ended questionnaires, are developed (and perhaps "field-tested") before the evaluation ever gets under way. Other parts, like case studies and participant observation schemes, are generally outlined in advance but largely developed or modified during implementation of the study.

The overall framework

In this chapter, we will deal principally with the quantitative side of the coin: how the underlying themes and questions of an evaluation are translated into indicators that guide data collection and that may yield important insight into the program under consideration. Qualitative procedures for investigating the themes and questions of evaluation research, which are typically more inductive and so rely to a greater extent on observation of actual events and phenomena rather than on collection of data to fit pre-defined categories, will be taken up explicitly in Chapters Eight and Nine. Their place in the overall "information generation matrix" of an evaluation will be highlighted in this chapter, however, and it should be clear throughout how these two inquiry strategies complement each other in evaluation design.

All the components of an evaluation design should therefore fit into the same general framework and contribute to elucidating the same underlying set of questions or themes. I do not mean here to deny the great worth of open-ended studies or the real possibility that some of the most valuable findings will concern largely unanticipated topics (quite the reverse: see the chapters on qualitative evaluation). Nor do I want to suggest imposing a straight-jacket on all components of an evaluation. In fact, the overall framework of an evaluation design can - and sometimes should - change form during its implementation.

But changing form is not the same as not having one. The point is simply that evaluators should always be clear about *why* any given study has been undertaken and at least what general areas of concern it is designed to elucidate. One of the classic problems in evaluation and research is collecting information without knowing how it is to be used and thus ending up with reams of material that simply gather dust on the shelves or (the more recent version) clutter up computer hard drives. Though a good analyst of secondary data can usually retrieve something of value from this sort of information dump, the fault really lies with evaluation design and with a failure to think through well ahead of time what data are being sought and why.

This point is analogous to the one made in Chapter Three about using "concentric" designs in participatory evaluation in order to ensure that what local program beneficiaries investigate or debate and what central evaluation personnel study converge around a common set of topics. The fundamental purpose is to preserve the possibility of ex-

change between the different levels of evaluation inquiry going on and to promote broad debate among program stakeholders concerning the meaning of results and concerning desirable future directions for the program.

In either case, the program model (context-input-process-output-outcome) may provide the broad framework for the evaluation. By specifying themes, topics and questions within this framework, one has begun to flesh out the design of the evaluation exercise at hand. The issue of interest in this chapter is how to move from such broad terms of reference to specific field studies that in fact furnish the information needed, and back from this data gathering to the broad questions and concerns that underlie the evaluation.

Crafting good indicators

It is never possible to collect all the information or to observe all the behaviors that might theoretically be required to answer the underlying questions of an evaluation, even if these themes have been carefully defined and explicated ahead of time. In order to know, for example, whether or not participants in a certain training program apply the agricultural techniques presented to them more or better than non-participants, one would practically need to be behind every one of them - and every member of an appropriate control group - during the entire duration of the growing season, something which is as undesirable as it is impossible.

Often it is difficult to observe or measure *directly* the phenomena which interest us because of their very nature. This is the case, for example, with respect to attitudes and intellectual competencies. While the age, height and weight of a person are all more or less obvious to an observer and can be determined with relative ease, attitudes and competencies are not visible to the naked eye. One can only try to determine them by induction from other more accessible observations, like visible behavior or the person's words and responses.

In both cases mentioned (impossibility of observing all relevant behaviors and "invisibility" of many important themes), the essential subjects that interest us are not entirely or at least not directly measurable. We can only try to define a limited number of phenomena that can be observed and that seem to us to indicate with some reliability the nature and intensity of the non-observed characteristics - barometers, so

to speak, of the invisible attributes. An "indicator" is therefore a phenomenon that one can observe or measure which informs us about another less accessible factor - in technical terms, a variable whose values are closely correlated with the nature or intensity of an unobserved phenomenon.

As an indicator of the degree of application by a group of trainees of an agricultural extension message regarding fertilizer use, for example, one might propose the proportion of group members who bought at least two sacks of fertilizer before the beginning of the growing season following the completion of their training. (Note that in order to define this indicator, it was necessary to better explicate the particular element of agricultural training in question. Moreover, adoption of such an indicator supposes that one can actually obtain information on the quantities of fertilizer bought by the former trainees. This example raises other problems to which we will return presently.) In a similar vein, in order to assess the attitudes of this same group of trainees regarding (for example) the benefits of agricultural marketing cooperatives, we might think of presenting them individually with a certain number of "propositions" concerning (e.g.) the value, uses and feasibility of cooperative schemes. Their reactions to these propositions might then be considered to constitute an indicator of their attitudes.

It is sometimes helpful to draw a distinction between *general indicators*, which define the type of information that will be used as a barometer, and the *operational indicators* - sometimes also called "indices" - which specify the exact and often fragmentary data which should be collected. In the case of the application of the agricultural extension message, the general indicator might be the proportion of trainees who buy fertilizer after the completion of their training. The operational indicator would be a good deal more precise and might be formulated in rather full-dress manner as follows:

The degree of application of the agricultural extension message on fertilizer will be indicated by the proportion of the group of trainees having completed training between June and December 1999 in X and Y districts who buy at least two sacks of fertilizer in any of the farm supply outlets of these districts by June 30th of the following year. This information will be obtained through examination of the sales records of the agricultural outlets in question and the records of the local agricultural extension agent.

The use of these supplementary details (or any intermediate level of specification that one may choose to adopt) is that they specify very clearly the path that leads from the evaluation theme in question, through the indicator and up to the instrument or data collection procedure that will need to be adopted. The clearer this linkage, the easier it will be to actually collect the data and then interpret them afterwards. In fact, *clear unifying threads of this type should link all stages in the conception and execution of the evaluation*. In particular, it is helpful to envisage quite early in the process the types of data analysis that will be undertaken once the collection stage is past, because these final critical convergent phases in the evaluation will be much easier and more fruitful if, by their form, the indicator scheme and instruments lend themselves to such interpretation and analysis.

Strong and weak indicators

Definition of indicators poses in very concrete fashion the problems of reliability, representativeness and (particularly) validity - in short, data quality - discussed in Chapter Five. It is obvious that in selecting one or two types of directly observable phenomena to gauge an outcome like the health improvement consequences or agricultural applications of a training program, one runs a real risk of missing the target or of collecting data which have such a tenuous relationship to the issues of concern that they provide nearly no insight.

In other words, each time one chooses a given piece of information to stand for another that defies direct observation or to represent a much broader theme that it is impossible to probe exhaustively, critical problems of validity arise. How well do the types of information selected actually "indicate" the aspect of the situation one wishes to understand?

- Are the results of a test of accounting knowledge administered to former cooperative management trainees really indicative of their practical competence?
- Do the points of view of the regional inspector and the community association president concerning the effectiveness of the local literacy instructor give a sound basis for assessing his work performance?
- Or would other information be necessary?

Finally, to return to the example cited above, to what extent does the proportion of a group of trainees who buy fertilizer from given outlets within a set period furnish a valid indicator of the degree of application of a related agricultural extension message? The data on fertilizer purchases can hide other factors that have little to do with application of extension messages. Some trainees may have either bought the fertilizer elsewhere or gotten theirs through third parties. Others may have bought it strictly for resale or to make use of it in ways that have little to do with increased agricultural production (as insecticide to mix with house paint, for example). In each of these cases (and a number of others), the information defined by the indicator would reveal little about the true proportion of the group that applied the extension message in question. It is questionable whether such information would be worth the time and expense involved in gathering it.

Indicators can thus be of good or bad quality, as well as every gradation in between. The quality of an indicator and the validity of the information that it furnishes are essentially synonymous, and both depend on the degree of congruence between the information given by the indicator and the different dimensions of the evaluation theme. Ideally, the indicator should inform us about all the aspects of the theme (or all of its components) and nothing else - an objective which is difficult to accomplish, but less difficult if the theme has been well defined.

It is very important to refine the indicators established for an evaluation exercise by a careful effort of analysis, reflection and discussion both among evaluators and with stakeholders themselves. The agricultural extension indicator cited earlier certainly requires some modification. If, despite the problems raised above, it is felt that data on fertilizer purchase do constitute the best available indicator of the degree of application of the extension message, one should at least try to take account of the possibilities of distortion that have been mentioned - by specifying, for example, a method for spotting and eliminating the purchases effected by or for third parties.

Often it is worthwhile to think of adopting two or more indicators for any given evaluation topic, since triangulation between these separate points will likely furnish a more valid appreciation of the situation than any single indicator could. In any and all cases, decisions made concerning the nature and the form of an indicator and concerning the way in which it will be translated into operational terms will depend almost

entirely on the specific situation being evaluated or researched. Tailoring good indicators therefore requires both a good understanding of the "evaluand" (the subject being evaluated) and a considerable measure of creativity.

If indicator definition has an immediate impact on the validity of evaluation data, it may also have an important effect on their reliability and sampling error as well. In the final analysis, the reliability of data is determined principally by the nature of the instrument used to collect them and the way in which this instrument is actually handled. Does the instrument as actually employed in the field enable evaluators to collect data that are relatively stable and accurate? At the same time, however, the quality of the instrument depends partly on measurement difficulties inherent in the indicators that have been chosen. If the data selected as indicators are very difficult to apprehend, development of an instrument will pose serious problems and results risk being quite unreliable. In addition, the representativity of data - which is principally determined by the sampling strategy effectively carried out in the field - will be affected by the "accessibility" of the information defined by the indicator. If the data in question are only available for a fraction of the population, for example, the sample of the study will obviously be distorted.

In other words, in defining indicators evaluators must concern themselves with the accessibility and measurability of the different data they propose as well as with their relevance to the themes of inquiry - that is, their validity. Choices in this domain are almost always trade-offs: one sacrifices a little reliability to have more pertinent information, or vice versa. Indicator definition thus requires a high level of creativity.

The example in Box 6.1 demonstrates in practical terms the issues and choices involved, and it introduces another distinction that may prove useful - one between direct, indirect and proxy indicators that expresses how closely the indicator in question approximates the actual phenomenon that we are concerned to assess. Remember that this distinction is often related to the validity of the indicator: the closer one comes to actual observation of the phenomenon of interest, the greater one's confidence that the data speak to the central issue at stake.

Qualitative and quantitative indicators

Indicators can be more qualitative or more quantitative in nature. An example will help to illustrate the point. Let's consider an evaluation

Box 6.1. Indicator choice in evaluation of Ghanaian literacy

The problem of indicator choice, the tradeoffs involved in opting for one type over another, and the methods by which different indicators of the same themes and variables are often combined in reaching conclusions, can be illustrated by an effort to evaluate the performance of the chairs of local literacy committees in northern Ghana. If one can define carefully what the chairpeople are expected to do in support of the literacy program, then it becomes possible to design a variety of indicators to assess the degree to which they are indeed performing these roles.

- "Direct indicators" would involve actual observation of the interested parties in the execution of their duties and notation of the frequency and completeness - and/or rating of the skill - with which they accomplish them.
- "Indirect indicators" would be represented by written records or other "artifacts" (committee minutes, evidence of visits to the center...) that help to establish what the chairperson has done and with what frequency he or she has done it, in the absence of direct observation.
- "Proxy indicators," on the other hand, might be evidence of the effects of the fulfillment or non-fulfillment of duties - like the top-notch or dilapidated state of the facilities, and the regularity of student participation - that give at least an impression of the chairperson's performance even if they cannot be as directly linked to actual measures of work accomplished.

In addition, there are a whole set of possible *qualitative* indicators, which include self-assessments by the committee chairs themselves, discussions with other knowledgeable parties, and detailed ("ethnographic") observation of an actual day in the life of the center...or of the chairperson him- or herself! The advantages of such more open-ended procedures in the latter case might include (a) the fact that they allow the evaluator both to take account of beneficial or non-beneficial things the chair does which are not in his or her official description of responsibilities and (b) that they put the notion of "performance" in social and personal context.

Choice of actual indicators to be used in such a case will always be a simultaneous function of the ease with which the information can be collected, the depth and accuracy of insight that it affords, and the resources and time at the evaluators' disposal. But it is fair to say that evaluators will usually choose to combine two or more procedures in order better to "triangulate" on the actual situation - often one procedure that is more quantitative and summary, another that is more qualitative and in-depth. The former might be represented in this case by data on performance of duties, the latter by interpretive interviews.

topic concerning the work performance of the instructors of a given training program.

To begin with, we need to explicate the theme itself more fully. What, for example, is meant by the "work performance" of these instructors? What do we really want to know? We could be talking about the degree to which they faithfully implement the prescribed curriculum, about their creativity in developing exercises and practical examples, about the clarity of their oral presentation of material, their facility at eliciting student participation, their faithfulness to professional duty, their degree of success in raising competence levels among their students, or a number of these elements at once. Let's suppose for the sake of argument that we wish to assess the instructors' *faithfulness to duty* or *professional reliability*.

Other specifications are immediately needed to clarify the meaning of these terms (for us) and the manner in which we propose to apprehend the behaviors in question. What constitutes "faithfulness" or "reliability"? What are the elements of this theme? One could focus on punctuality and regularity of attendance, time spent in the preparation of lessons, constancy in the correction of student exercises, etc. All these behaviors might indicate an instructor's "faithfulness" to his or her assigned duties. How then would we go about assessing these different characteristics of instructors' behaviors? What data would be collected and how? What *indicators* should be adopted? Some aspects of this theme lend themselves easily to quantitative assessment. If the instructor (or someone else) keeps an attendance register, it should be possible to determine at least approximately the number of class sessions held in the center and to compare this to the number officially envisaged by the sponsoring agency or by local authorities. Such a proportion would constitute a quantitative indicator of instructor performance. (For the purposes of the present discussion, we will presume that the data recorded in the register are reliable, though this is frequently not the case. In practical applications, some cross-checks between the register, the commentaries of trainees and local authorities and the monitoring forms filled out by regional staff should enable the evaluator to get a better idea of the actual reliability of the data.) One could also get roughly comparable data on the manner in which each instructor prepares his or her lessons and corrects student notebooks. To quantify these appreciations, the evaluator might develop a system of categories for assessing the quality of the work in relatively standard fashion, beginning with careful

analysis of a certain number of notebooks and lesson preparation books in a sample of centers in order to identify the different types or levels of lesson preparation and notebook correction performance most frequently encountered. Since this procedure brings into play the judgment of an outside observer - namely, the person responsible for devising the categories, who will ideally have been trained for this sort of work - it can be marginally less reliable than the simple calculation of a rate of attendance, but on the other hand it furnishes information that is typically richer and more suggestive.

The evaluator could equally well pose questions concerning the instructors' performance directly to other persons who know each of them and their work - for example, regional literacy staff, local literacy committee members, or trainees - either to obtain information on one or another of the components of their duties that has been defined or quite simply to solicit these observers' own open-ended evaluation of each instructor's work performance. The observations and assessments offered by these informants would then constitute a *qualitative* indicator of the evaluation theme, one naturally subject to certain distortions linked to the subjective nature of the responses but which at the same time may offer greater insight into the issue. And we shouldn't forget the method of self-assessment which would entail asking the instructors themselves to evaluate their own work performance, to explain on what basis they make this assessment, and to comment on the factors that explain it. If their answers are by definition subjective ones, in some sense no one is better placed than the instructor him- or herself to "observe" the behaviors in question and to draw practical conclusions from these observations - and certainly no one is better placed to put the conclusions of such an evaluation into practice!

Several lessons can be drawn from this example of indicator choice. First, it should be evident that one can imagine a great number of indicators for any particular evaluation theme, and that some will be more valid and easier to use than others. A good indicator should have a clear logical link with the evaluation theme; it should prescribe data that are relatively easy to collect with the instruments or personnel at the disposal of evaluation staff; and it should be determinable in a relatively stable and precise manner. Developing indicators which answer to these three criteria - logical relevance (i.e. validity), feasibility, and reliability - is never a simple affair.

Second let's note that indicators may be situated at different spots along a scale that extends from purely quantitative measures to entirely qualitative ones. The distinction between the two poles is largely a matter of the degree to which the subjective or unstandardized judgments and observations of the evaluators, of the "valuees" or of third parties are included. A quantitative indicator - like the number of class sessions held as a proportion of those initially envisaged - only entails the participation of the evaluator in the field under standard and well-defined conditions. One has simply to collect a specified kind of data and plug it into a specified proportion. Systematic examination of a series of documents in order to draw from them observations about the behavior of the instructor requires already a heavier dosage of researcher intervention and interpretation, but subjective elements can be minimized by the standardization of categories - that is, one can define ahead of time the characteristics of these documents and the classification grid to use. Such an indicator is situated somewhere between the quantitative and qualitative poles on the scale.

Nearer the other end of this spectrum are found the open-ended assessments and responses of persons who are not part of the evaluation team - i.e. unstandardized observations, whose richness generally lies in the fact that they have not been squeezed into a uniform set of categories. These data are more difficult to transcribe but typically contain much more information than closed-ended responses. Such questions are not always as open-ended as they first appear, of course. Answers can be biased, consciously or unconsciously, by the way in which (and the circumstances under which) the question is posed; and evaluators need to be careful about creating signals and situations which suggest acceptable answers to their respondents, or which provoke strong reactions that may color the information provided. But the net is at least potentially cast wider in open-ended formats than in closed-ended ones.

Strengths and weaknesses

Each type of indicator has its strong and weak points. Quantitative indicators can be very precise and reliable. In addition, they are relatively easy to aggregate and compile; and, since the resulting data lend themselves to statistical analysis, one can easily put together a picture of the situation across an entire zone and investigate associations between this factor and others (such as, for example, the prior training of the instructors in question) through methods of cross-tabulation and correlation.

However, quantitative indicators are often weak on validity. The reasonings and derivations leading from an evaluation theme to an indicator that can be easily quantified and measured are sometimes so tenuous or torturous that one can legitimately ask if the data obtained in this manner say anything at all about what initially interested us. In addition, quantitative indicators typically furnish less information than qualitative ones. In the attempt to describe a situation in numeric terms, one is forced to eliminate most of its particularities and to retain only the categories and degrees of manifestation that permit easy aggregation and cross-situational comparison. The enumeration of the number of instructors who prepare their lessons "well", "moderately well" and "poorly", for example, says very little about their actual practices or about the real quality of work performance that those practices reveal - let alone about the factors that contribute to such variations in performance.

Qualitative indicators, on the other hand, can furnish much richer and more subtle appreciations of the theme in question as well as data that are often more immediately relevant. Rather than decomposing the theme into numerable component behaviors or facts, one has recourse to people (the evaluators themselves or well-placed informants) who have observed the phenomena in question in context and in much of their complexity, and who may therefore be able both to speak more directly to the theme in question and to convey a richer sense of what actually happens than one could derive from numbers alone. To the degree that the informant understands the phenomenon and the inquiry theme, the data that he or she furnishes will typically be both richer in detail and more valid (i.e., more directly relevant to the issue of concern) than purely quantitative information.

But the flip side of the coin is that these data are much more difficult to compile and aggregate or to compare from one region to another than quantitative data, and they often pose more serious reliability problems. Observations and assessments of a given phenomenon gotten from one type of informant at one point in time may not coincide with those gotten from another at a later point, since the particular perspective of the individuals concerned and the events surrounding its expression enter into the data. Part of this is fine: such variations allow one to take account of multiple viewpoints. Part is not: it is hard to summarize handily these idiosyncrasies and to separate out objective and subjective elements.

Multiple indicators

The example used above brings out a third point: it is entirely legitimate and often very helpful to use *several different indicators* - some more quantitative and others more qualitative - in order to study a given evaluation theme, because the triangulation of these varying points of view and differing levels of detail should provide more insight than any single type of data could. Quantitative and qualitative indicators are thus natural complements. For example, one obviously gets a better appreciation of the knowledge levels of a group of students by comparing quantitative information on their test scores with qualitative data from interviews with them and their teachers than one would by relying on the test scores alone.

At the qualitative end of the spectrum, of course, indicators fade into the still more open and naturalistic type of study that is characteristic of qualitative research but may also be part and parcel of an evaluation design. We will speak more directly of the application of qualitative research methods to evaluation in Chapters Eight and Nine. Suffice it to say for the present that in the numerous areas where we really do not know how best to break our questions and concerns down into indicators or where we want to open our understanding to new possibilities and to unanticipated stakeholder points of view, the appropriate strategy is a highly qualitative one in which maximum attention is given to identifying particularly knowledgeable persons or very information-rich cases, then to gaining entry and acceptance in these settings and allowing what one learns from such an immersion to define progressively the exact questions, concepts and indicators that will unlock understanding of what is going on.

Output and process indicators

Since indicators may be devised for nearly any of the phenomena that characterize a program, it is obvious that they may have to do with very different aspects of it. Pertinently, to use the language developed in Chapter Four on systems thinking, one may devise separate indicators of context, of program input, of program process, output or outcomes. These distinctions themselves are not absolute but may be relative to the purposes of the evaluation. For example, whereas data on services and material actually delivered to the field frequently constitute output

indicators for an administrative evaluation (that is, as further described in Chapter Ten, an assessment of the operations and results of the organizations responsible for staffing and supporting the program), in an evaluation of *instructional* activities at the local level or in a more general assessment of the entire system these would constitute indicators of input.

The distinction most frequently made in current literature on both formal and nonformal education is the one between process indicators on the one hand and output or outcome indicators on the other. The charge often leveled at educational programs and their evaluation services is that they mistake the one for the other, or that they focus on input and process indicators like numbers of students and textbooks or average daily attendance and seldom furnish indicators of the "outputs" which educational effort is supposed to produce and which give a better sense of the effectiveness of the programs in question: namely, decent quality data on what participants have actually *learned* and what they have *done* with this new knowledge. This imbalance is not altogether surprising: reliable output or outcome data are both more difficult to produce and potentially more contentious than reports on attendance and teacher qualifications. (True learning and genuine application of lessons are more difficult to ensure than input delivery or attendance, and sometimes less dependent on the sole efforts of program staff; and people would naturally rather be held accountable for what can be relatively easily produced or what lies entirely within their control than for things that are more arduous, unpredictable or contingent.) In any case, at the Jomtien Conference on Education for All particular emphasis was given to the importance of developing output indicators for educational programs.

This is well and good, but obviously has its limitations as well. Recently in Florida, as in many other regions of the United States, central education officials have given great prominence to a school restructuring initiative, in this case entitled "Blueprint 2000," which is designed to decentralize educational decision-making while simultaneously holding schools accountable for the achievement of a certain number of goals adopted at the State level. The stated purpose of the exercise is school improvement. The seven education goals defined in Florida's Blueprint 2000 place particular emphasis on output measures of student achievement, graduation and readiness for further training and employment. The idea conveyed is that we are "at last" getting serious

about school performance, about what educational programs and personnel do and do not *produce*.

Learning is certainly the fundamental objective of educational effort, and it is understandable and legitimate that those who fund educational programs - the taxpaying citizen, in the final analysis in the North American case - should wish to know how effectively their resources are being used. But insofar as school improvement is in fact the central purpose, evaluators should be quick to point out that output data, whatever their accuracy and relevance, are not enough.

The key question for formative evaluation is not simply what results have been produced but *why* and *how*. If we don't understand what led to a given set of outcomes, we won't know how to improve them. In fact, unilateral fixation on output measures may have the contradictory effect of reducing the effectiveness of evaluation - both because one omits in this way some of the evidence needed to analyze the dynamics of the program, and, still more simply, because staff who believe they will be judged on the magnitude of results they have produced without any account being taken of the adequacy of resources they had to work with, or any recognition of other extenuating circumstances, are generally very loathe to cooperate or provide decent data. Significantly enough, such has been particularly the case for *adult and nonformal education* programs in Florida. Results of these efforts have been sparse, but resources have been sparser still - a typical problem (typical for adult education) of trying to make bricks with little if any straw. Under these circumstances, local staff have demonstrated a virtually allergic reaction to the idea of being evaluated purely and simply on the basis of output or performance indicators. They understandably feel that it could be suicidal.

So the distinction between output and process indicators is often a political as well as a technical one, and it is a possible bone of contention between those who fund programs and those who carry them out. Both types of data are unquestionably necessary to a decent evaluation, but evaluators need to be sensitive to their "surplus meanings."

Sources of information: Relating MIS and evaluation

Where do the data defined by individual indicators, and by the indicator system as a whole, come from? Since we are talking here about evaluation practice, it would be natural to suppose that they are all collected in the course of different evaluation studies. But that is not necessarily true,

unless we interpret "evaluation studies" in a very broad fashion. In fact, any periodic evaluation exercise depends to a considerable extent on data already gathered - as well as on personal impressions already formed - prior to the commencement of the evaluation itself: information in files, in notebooks, and in people's heads.

Any organization collects a certain amount of information on its own operations in the natural course of program implementation and administration. The importance of this regular "monitoring" has become very evident over the last 30 years as social and educational programs in all areas of the world have acquired increased experience with the implementation process and have confronted the difficulty of performing decent formative evaluation without a baseline of data on their daily operations.

The way in which monitoring data are regularly gathered as part and parcel of program administration and then fed back into the system in order to keep staff at all levels abreast of the evolution of activities has been described as a program's or organization's "management information system" (MIS) (cf. Chapman 1993). Businesses, public agencies, schools, nonprofit organizations, and associations of many different types all need some form of MIS to guide day-by-day decision-making. The term MIS may of course cover systems of very different natures and of very different levels of complexity or effective integration; and it does not necessarily imply or entail the use of computer equipment. "MIS" is, after all, simply a term that we have made up to group together the diverse monitoring practices of organizations and call attention to their information facets and to the benefits to be gained from better developing these. A practice of recording class attendance and achievement levels in notebooks at the training center level and then periodically sending totals from these documents to central agencies constitutes an MIS every bit as much as a highly computerized system for tracking multiple indicators of program quality. It is simply a more rudimentary one, and not necessarily a worse one. "Worse" and "better" have finally to do with the uses to which data are put and with the ratio of benefits derived from their collection and use to the costs in effort and resources required to do so.

Management information systems - or at least discussions about them - have recently come into fashion in educational institutions worldwide, and no self-respecting program would admit to being without one. Problems arise, however, when management information is

perceived as a concern and domain entirely separate from evaluation: separate staffs are created and separate instruments and procedures developed. Though there may be very valid reasons in large programs or organizations for harboring staff specialized in one or the other of these two related domains, it is critically important that their close relationship be understood and take precedence over conflicts of turf. In point of fact, there is a series or spectrum of information functions within any program that stretches from daily monitoring of activities through formative - and preferably participatory - evaluation of program implementation to end-of-period/summative evaluation of impacts and on to the various kinds of continuing applied research and R&D undertakings that furnish staff with new tools and intervention methods, as well as deeper insight into the social dynamics of their programs. All these information generation efforts should complement each other, the differences among them having to do mostly with their frequency and scope.

An information generation matrix

Indicator data may therefore come from existing information gathered for monitoring, MIS or research purposes, or they may be collected explicitly for use in the evaluation study at hand. To clarify these linkages and keep in focus both the overall framework of the evaluation and its essential complementarity with other data gathering activities of the program, I find it useful to draw up what might be called an "information generation matrix" - that is, a large table where one may picture the logical flow from evaluation themes through indicator definition to actual collection and analysis of data and which situates the various themes under investigation with respect to an overall program model. This is only one variety of an indicator "system", a topic of increasing interest to educational planners in recent years (cf. Chapman 1993; Easton et al 1991; Windham 1990).

An example of an information generation matrix (or a portion of one) is presented on Table 6.1 hereafter. The first column on the left concerns the general realm of the program model in question - context, inputs, processes, immediate results (outputs) or longer-term effects (outcomes). Further distinctions could of course be made within these in the manner indicated in Chapter Four. Subsystems in each realm might be distinguished, like teacher recruitment and training within adminis

Table 6.1. Information generation matrix: Instructional domain

DOMAIN	THEME	QUESTION	INDICATOR	INSTRUMENT	PROCEDURES		
					Sample	Data Collection	Data Analysis
CONTEXT	Math and accounting systems used by local merchants	Mental calculation strategies and capacities of local merchants	Qual: Observation of transactions in market and open-ended interview with sample of local merchants.	Research guide for indigenous math skills study.	Snowball sample in experimental center localities	By university personnel in July	Content analysis by university personnel
INPUT	Instructor characteristics	Training or apprenticeship undergone	Quant: Average nb. of days of training undergone by interviewed instructors Qual: Results of open-ended interview about how instructor learned job.	Instructor survey item #12	Stratified sample A	By auxiliary surveyors in May	During regional evaluation workshops.
				Instructor interview guide page 2.	Stratified sample B	By regional inspectors in Jun.	
PROCESS	Teaching quality	Participatory pedagogy	Qual: Results of evaluator participant observation of instructional session	Observation protocol page 4	Stratified sample B	By evaluation staff in Jun	Same
	Administrative efficiency	Organization of instructional material	Quant: Type of inventory system used for materials.	Instructor self-report form	All centers	By instructors in May	Same
OUTPUT	Knowledge and skill acquisition	Mathematical competencies	Quant: Results of validated self-evaluation of trainee competency on first year's lessons	Trainee self-evaluation booklet & summary sheet	Stratified sample A	By instructors in June, validated by inspectors	By regional inspectors for submission to workshop
OUTCOME	Literacy retention	Use of reading skills in everyday activities	Qual: Results of guided interview w/ sample of former trainees regarding current use of skills.	Former trainee interview guide page 3.	Experimental centers	By university personnel in July	Primary analysis and write-up by university.

124

Indicators and insight

trative (input) concerns, or the quality of visual aids in the instructional (or process) domain.

The next column (second from the left) specifies the focal themes chosen for the evaluation. These can be further discriminated into sub-themes in the third column from the left. Next come the indicators developed as means of investigating each theme or sub-theme. Two types of indicators - qualitative and quantitative - are illustrated. With more space, one could additionally detail both the general indicator and the "operational" ones that specify more exactly the actual data to be collected. The following column then defines the type of instrument or procedure - and potentially the item or component within it - that will be used for collecting the data in question. The last three columns on the right, finally, provide space for laying out some of the mechanics of data collection and analysis.

There is nothing sacrosanct or copyrighted about this table and the related procedures. They can be modified or simplified (or, for that matter, discarded) in whatever way one sees fit. I myself am not sure to use this instrument in the same form next time. But the underlying utility of such an exercise is that it forces evaluators to do several things. It impels them to -

- trace out clearly the linkages between the underlying themes or questions of the evaluation and the actual data that are collected;
- think out how the information gathered will be analyzed;
- picture the relationship and essential unity among the different elements and components of an evaluation; and, particularly,
- map the relationships among MIS data, evaluation data and related research information.

This last point is a critical one. I remarked above on the unfortunate separation and even conflict that may arise between MIS staff, evaluation staff, and research staff, despite the fact that all are contributing to an essentially unitary information function. One major advantage of the kind of scheme portrayed in Table 6.1 is that it graphically portrays the interwoven and complementary nature of these different functions, and does so within a framework that relates them all back to improvements in the program model and in its effectiveness. In addition, a scheme like this provides a sort of "road map" for using the data actually gathered to answer the fundamental questions or elucidate the priority themes

underlying the evaluation - and so for avoiding the classic pitfall in evaluation and research: collecting reams of data and not knowing what to do with them. One can always trace back the reasoning that led from such underlying questions to the specification of actual data to be collected in order to relate the information gathered to the central concerns of evaluators and decision-makers. This is useful not only in data analysis but equally in constituting a "paper trail" that justifies the design decisions made by evaluators.

Evaluation "mosaics" and decentralized MIS

This vision of a grand "information generation matrix" guiding indicator development and overall evaluation inquiry, though certainly helpful from a conceptual point of view, needs to be moderated a bit by recalling what was said in Chapter Two of this book about decentralized evaluation designs. Let's recall that nonformal education initiatives are increasingly localized and that the efforts of national agencies are supplemented and even overshadowed in numerous ways by those of NGOs and the other new actors in "civil society." It was suggested in Chapter Two that this is in fact a *favorable* situation for evaluation, both because it serves to increase the variability in findings on which productive data analysis depends and because it potentially means a great increase in the number of those actually doing evaluation work. But it also has consequences for the way in which evaluation design and such things as grand matrix schemes are conceived and used.

As far as indicator schemes like that pictured in Table 6.1 are concerned, decentralization may mean that these tools serve as a suggestive framework for sharing evaluation tasks among organizations and undertaking complementary work rather than as the sovereign "game plan" for a single evaluation unit. So much the better! The fundamental objective, in any case, is not to impose a lock-step design, but instead to stimulate thought about the aspects of the program (or set of associated programs) that it is most important to assess and about the type of data that will help everyone to do so. Quite possibly, matrices like the one pictured will be of more use as a training tool or a discussion theme than as an actual implement for evaluation design.

A similar point could be made with respect to management information systems. Though the popular image of an MIS involves, not

without reason, a large computerized system for ingesting data on the many tentacular branches of an administration or an organization, in fact decentralized units have no less need of good information on their own operations than do centralized ones, and they may be able to put information to more immediate and more productive use. It is true that the nature of information needs changes in subtle but important ways as one moves from center to locality, where the concern is less with distant monitoring and compliance control than with proximate feedback and improvement. But those concerns are arguably much more in harmony with the supposed purposes of MIS anyway. It could even be claimed that vast national management information systems are something of an aberration, insofar as data warehoused in this manner seldom serve more than abstract statistical purposes. In any case, under increasingly decentralized regimes the challenge is to work out procedures and mechanisms for *local* management information systems. If these are providing valued data and grist for reflection to staff on site, they will also be able to furnish all that is required for central coordination and management. On the contrary, centralized MIS's that depend on branch staff collecting and passing up the line monitoring data that they never have reason to use themselves risk seeing the reliability and validity of their inputs take a rapid nose-dive. Reliable information on field operations is essentially a by-product of the activity of local organizations that are collecting and using data for their own purposes.

Conclusions

"Indicator systems" have been very much the rage of late in educational circles. Accountability pressures on schools and training programs in many countries have caused a premium to be put on the design of compact sets of data that provide a quick overview of the state of programs, both for decision purposes and for public relations ones. In addition, and not entirely coincidentally, such schematized forms of quantitative data are those that best lend themselves to incorporation into centrally computerized management information systems as well as to aggregate analysis by parties having little familiarity with the particulars of each program and location.

The fact that indicators must be both qualitative *and* quantitative if they are to help create the sort of insight that will contribute to improv-

ing educational performance is often lost in the shuffle. Such "triangulation" of measurement, observation, informed judgment and subjective feeling is the very heart of good decision-making but is more difficult (though by no means impossible) to assimilate into computer format. Analysis of carefully-chosen and carefully-collected quantitative indices of program performance is certainly one important part of this gestalt, but only one. Happily, the increased miniaturization, shrinking cost and broader dissemination of computer technology make it at the same time increasingly feasible for local programs to create their own "management information systems"; and at this level it is perhaps more likely that such quantitative data will be appropriately combined with other types of information in the decision-making process. Creation of *local* management information systems is in fact the logical objective of MIS programs, though this point is often forgotten. In fact, with an eye to the case that we have been making throughout this book about the importance of participatory evaluation and of stakeholder collaboration in technical work, it seems fair to say that the real concern in MIS should be progressively to transform the Management Information System into an "Informed Self-Managing System"¹ - that is, a set of procedures with appropriate computer support that enables local stakeholders to put together quantitative and qualitative information in ways yielding the insight they need in order to guide their own programs.

But whatever the current state of the MIS debate, the real importance of indicator systems that we have tried to emphasize here is in fact prior to such considerations. It quite simply involves building up and improving the linkage between the underlying questions that drive our evaluations on the one hand and specification of the actual data - both qualitative and quantitative - that we gather in the field on the other. Time spent trying to understand and develop that linkage is never wasted effort.

Notes

¹ Or, for short, "IMS"? We may have a winner here! As this book amply demonstrates, evaluators are not immune to the contagion of acronyms. There is probably no major damage done, provided one doesn't let them get the upper hand.

References

Chapman, D. 1993. *From Data to Action: Information Systems in Educational Planning*. London: Pergamon Press.

Easton, P. and Dwight Holmes, Howard Williams and Joy DuPlesses. 1991. *Efficiency and the Collaborative Design of Educational Indicator Systems*. Technical monograph. Tallahassee, Fl: Learning Systems Institute, Florida State University.

Windham, D.M. 1990. *Indicators of Educational Effectiveness and Efficiency*. Tallahassee, Fl.: IEES, Educational Efficiency Clearinghouse, Learning Systems Institute, Florida State University.

CHAPTER 7

SAMPLING STRATEGIES

Sampling is an imperative part of any evaluation design, even if - a rare occurrence indeed - resources allow very extensive and extended data collection. One can never gather fully detailed information on *all* the aspects of a program of interest to an evaluation nor on all its participants. Evaluation and inquiry effort are not free: it obviously costs something in time, energy and resources to collect and analyze data, and every additional dimension or study that is added to an evaluation design increases at the same time the expense involved in implementing that design. Cost-benefit considerations necessarily enter into play.

Moreover, an exhaustive approach to evaluation and to data gathering would not necessarily be desirable even if it were possible. Suppose one had the resources at one's disposal to ensure thorough monitoring and data gathering on every facet of a program. Such a police-state and tight surveillance version of evaluation practice would likely provoke such negative reactions that it would soon have to be abandoned. In short, evaluators are invariably confronted with a sampling problem: how to identify and select the particular studies that will be conducted, the particular methods that will be used, the particular locations that will be visited, the particular activities that will be observed or assessed, the particular program beneficiaries and stakeholders that will be reached, and the particular times at which the inquiry will take place - given that one can, in any case, only follow up on a minority of the possibilities in each of these domains?

This selection problem is also a problem of inference. Since it will only be possible to collect data on a minority of the facets and beneficiaries of a program and at a few particular points in time, what can one do to ensure that the information collected in this way yields the best possible picture of the overall program? In other words, given scarce resources, how does one optimize the "yield" of evaluation studies? How does one get the most "insight" for cost? In an era of budgetary austerity and agency accountability, it is just as important for evaluators to be able to justify their decisions about resource use as it is for any other branch

of the organization. Sampling problems therefore arise wherever one turns.

The varied dimensions of sampling

Sampling is not just an issue of selecting potential human respondents or physical locations from a broader population. Rather, it is a question of making limited inquiry and monitoring capacities best serve the purposes of program improvement. There are in fact at least five distinct though interrelated kinds of sampling; or five domains in which sampling may need to be performed:

- *Demographic or geographic sampling* - the type popularly associated with the term "sampling" - concerns techniques for choosing subjects from among a population of people, sites and/or locations.
- *Temporal sampling*, which is less widely recognized but no less important, involves methods for selecting the particular times or junctures at which given aspects of a program will be evaluated. Programs operate in time, both cyclical and secular - that is, they generally both repeat a cycle of activities and evolve from initial to concluding phases. Data gathered and impressions derived at one point in time will not necessarily be the same as those that one would have obtained at other points. There are consequently important "sampling" decisions to be made in this realm as well.
- *Thematic sampling* concerns the choices made about what dimensions of a program to scrutinize. Should one be focusing particularly on context, input, process, outputs or outcomes (impact)? And within each domain, what are the key issues to examine? There are, in any case, many more potential topics than one can realistically investigate in any given evaluation exercise.
- *Theoretical sampling* refers to the choice of conceptual or disciplinary framework to be applied to the study. Will the concepts be drawn most heavily from economics, from sociology, from political science, from anthropology, from educational psychology or from other domains? What mix of these approaches will prevail? How will "program theory" be analyzed and assessed?

- Finally, *methodological sampling* concerns choices made about the particular evaluation methods to be adopted - i.e., the blend or preponderance of qualitative as opposed to quantitative approaches, the specific types of studies to be undertaken, the analytic uses to which data will be put, and so forth.

The point made here about the variety and breadth of sampling issues reinforces the one emphasized at the end of Chapter Five on data quality: the issue of selection among alternatives is a fundamental theme of evaluative inquiry, and concerns about reliability and validity are also, in a sense, sampling problems. The challenge of selecting from among multiple possibilities - what the French call the "confusion of choice" (*l'embarras du choix*) - is a good one to have, because it is careful consideration of alternatives that leads to good design, just as it is comparison shopping that generally leads to good purchasing decisions. In this chapter, we will make explicit the broad choices available to evaluators in the area of sampling methodology and offer some guidelines for the selection process.

In the pages to follow, we will talk about quantitative and qualitative varieties of sampling methodology and about the sort of macro-sampling issues involved in evaluation design choices. Greatest attention will be given to different approaches to demographic and geographic sampling, as that is the aspect of the methodology that has been the most fully developed. We will return to other aspects of sampling at the end of the chapter. In the realm of sampling, as in so many other dimensions of evaluation methodology, quantitative and qualitative approaches are natural complements. The common denominator that links them is the possibility of systematic selection among multiple potential options. The opposite of both is haphazard or convenience sampling, about which we will have something to say in conclusion.

Quantitative sampling

Quantitative sampling methodology is designed to ensure that the units or members of the group selected for study are as statistically representative of the broader population of concern as possible. It is largely, or at least "normatively," based on theories of probability - that is, applica-

tions of probability theories establish what the methodology *should* look like as well as a criterion for determining how far one is from that ideal, and statisticians make a clear distinction between "probabilistic" and "nonprobabilistic" ways of going about sampling.¹

Probability samples are basically those in which the final units are selected by some sort of chance procedure - whether flipping a coin, using a table of random numbers, or drawing numbered balls from an urn - and in which probabilities of selection are known: that is, one can estimate in advance the likelihood that a unit with a given characteristic will be selected on a random draw. Probability theory offers a means for making this estimate, *provided that the characteristic in question is normally distributed in the population* and that its variability can be at least approximately assessed. Normal distributions are, roughly speaking, those patterns frequently occurring in nature that result from chance apportionment. In "bimodal" distributions (like the apportionment of newborn babies between the male and female sexes), one expects approximately equal incidence of both categories or values of the variable. In the case of characteristics with multiple levels or continuous gradations, most units are grouped around the medium values and the frequencies tail off at either end of the distribution (high or low) in symmetrical manner. If the characteristics of interest to the evaluation or survey are normally distributed in the population and one can at least estimate their variability, then probability theory provides formulas and means for determining the chance that a given characteristic will show up in a given selection as well as for assessing the number of units that one would need to draw in order to have a representative picture of the population as a whole with respect to these characteristics (or, more precisely, a picture accurate within some specified margin of error).

"Nonprobability sampling methods" is the term statisticians use for selection procedures that do not involve the chance element and to which probability theory can therefore not be applied. They include haphazard samples (the units one happened to find or run across), quota samples (composed by adhering to guidelines about representation of groups defined by nonprobability criteria), and elite samples (people chosen because of their special knowledge) among others. Qualitative sampling - to be more thoroughly discussed below - could be said to fall into this category. For some statisticians and researchers, the term "nonprobability sample" carries a negative connotation: the implication

is that the procedure has been insufficiently rigorous to provide a basis for estimating margins of error and that any conclusions drawn from study of such a sample would not be "scientifically" justified. But this is manifestly misguided. Many characteristics of interest to evaluators and researchers (e.g., personal or family wealth) are *not* normally distributed in the population, and probability theory is of little help in selecting samples to investigate them unless one is first able to understand the varied dimensions of the situation and categorize the population into more homogenous groups.

Probabilistic methods also are not of much assistance when one is looking for cases that are particularly illustrative of some problem or for "key informants" specially qualified to give insight into some situation. This doesn't mean that the evaluator is excused from adopting a systematic procedure for selecting samples. The qualitative methods discussed further along in this chapter can be quite rigorous for the purposes they target. It simply means that the criterion of good sampling should be whether or not one has respected some type of systematic procedure adapted to the purposes of the evaluation, not whether or not one has drawn numbers from a hat or employed a table of random numbers. Haphazard samples would not satisfy this criterion but many kinds of qualitative samples would.

Entire books have been devoted to the subject of sampling (e.g. Morris 1988). Given space limitations, the most useful approach is to get at some of the underlying notions in the area, which are too often passed over in the rush to compute sample sizes -- and to get a sense of what is and is not worth doing. These are the objectives of the present chapter.

Some terminology

Before proceeding further, we need to define some words. In the sampling lexicon, the term *population* refers to whatever entire group one wishes to characterize or obtain information about. In a study of continuing learning behaviors among school dropouts, the population of interest might be all former students of no more than 20 years of age who had attended school at least through the end of the sixth year and then discontinued their formal schooling at some point thereafter. Obviously, one must take some care in *defining* the population of any given study or

evaluation - that is, the exact dimensions of the group to which one would like to be able to "generalize" the results of the study.

Implicit here is a second specification of importance: the *units of analysis* of which the population is composed. We could be talking of a population of former students, a population of teachers, or a population of schools or places. In the former two cases, one would sample individual students or teachers, whereas in the latter two one would be sampling institutions or localities. In some evaluations or for some evaluation purposes, it may be more important to examine typical process or average performance by training center or school than to carry on a full-fledged inquiry at the individual level. The decision concerning level of inquiry and unit of analysis needs to be carefully considered and thoughtfully made in advance. Failure to do so can lead to what researchers often term the "ecological fallacy" - namely, sampling at one level of analysis and drawing conclusions about another. Such a problem might arise, for example, if one collected information from a sample of schools and then used it to generalize about or characterize the attributes of individual students within those schools.

Finally, the *sample* is the group of members or "units" from this population on or with which the study will be conducted. To say that a sample is "statistically representative" of a given population means that one has the same chance of encountering any given characteristic in the sample as one does of encountering it in the population at large, and thus that the *proportions* of units with different characteristics - females as opposed to males, literacy instructors with a secondary education diploma as compared to those with other levels of formal education, training centers equipped with electricity as compared to those without it, etc. - is the same in the sample as it is in the overall population.

We should, of course, say "approximately" the same, because the characteristics of a sample never coincide exactly with those of the population at large unless all members of the population are included in the sample. As we saw in Chapter Five, there is always a margin of error in inquiry data; and the purpose of sampling techniques is, on the one hand, to reduce that error as much as possible and, on the other, to get an idea of how large it still is and how much caution we should still take in generalizing from our findings.

In the quantitative sampling tradition, three things determine how representative any given sample can be: (a) the variability of the charac-

teristic(s) of the population that one is trying to assess; (b) the size of the sample; and (c) the method actually used for establishing that sample.. The latter two factors - sampling size and sampling method - are potentially under the control of the evaluator. The first factor (variability of the characteristic in the population) obviously is not; but it is so important and so often overlooked that we need to spend a moment on it first. Sample size and sampling method will each be considered in turn. Though these considerations only apply strictly speaking to quantitative sampling, they provide insight at the same time into the general problem of choosing small groups representative of larger populations of interest.

Attribute variability

Let's take a far-fetched example. Suppose that a reconnaissance mission is sent to the earth from the planet Mars to determine the feasibility of an invasion. The Martians have never before set foot on earth, but intelligence gathered from space ship overflights (referred to in earth terms as "UFOs") has informed them that the earth is inhabited by animate reasoning beings. Now Martians are technologically advanced creatures but rather small in stature. Their infra-red space probes have detected living forms of various kinds on earth, some as tall as 80 meters (probably trees in a tropical rainforest or redwood grove, though they didn't know it). They believe on the basis of evolutionary science that creatures of a single or related species that developed together are likely to be of roughly similar stature - at least within an order of magnitude, as we learned to say in Chapter Five. The question that the reconnaissance mission is supposed to answer through observation of a sample of human beings is simply whether these intelligent earth creatures are more than fifty meters tall or in fact only a meter or two in height. How big of a sample do they need in order to draw accurate conclusions?

The answer of course is ... one. A sample of one single person from a population of several billions will be enough to answer their question. The reason is that variability in the population is virtually nil, at least in terms of the way in which the variable is scaled. For their inquiry, the variable "height" is a categorical one with two values: (a) more than 50 meters tall; and (b) less than 50 meters tall. The whole human population of the earth falls in category (b) and no one outside of fairy tales falls in

category (a). With so little variability, it doesn't take much of a sample size to answer the question.²

The example is purely fanciful but the point is valid: the greater the variability in the population of the characteristic or characteristics that you are trying to assess, the larger a sample you will need (all other things remaining equal) in order to "capture" this variability with any degree of certainty - that is, in order to ensure that all major or relevant subgroups of the population are represented in the sample that you have assembled and that you can therefore accurately characterize the composition of the population in this regard.

That is all well and good, but there remain at least two annoying flies in the ointment. First, how do you know the variability of the characteristic of interest ahead of time? Isn't that generally one of the things that the inquiry is supposed to discover? By adding enough provisos and conditions to our Martian example, we can suppose that this reconnaissance party already had a good sense that humans were either all over, or all under, 50 meters in height. But one doesn't have such liberty in the real world, and evaluators often do not know these variabilities with accuracy ahead of time. The answer is simply that half a loaf is better than none. If one can at least estimate roughly the variability in question - perhaps on the basis of the demonstrated variability of some known factor that is correlated with the characteristics of interest - then there is a basis for establishing sample size. If we are interested in average household income in some community and at least know in advance something about the variability in educational attainment of its members - as well as something about the relationship between education and household income in that society - then we may have a basis for estimating the magnitude of income variability. At least we probably are in better shape than someone who had none of this information.

The second problem with this approach is that a survey or evaluation is typically concerned with many different factors and variables, rather than one single population characteristic. Which one or ones of these factors should serve as a basis for estimations of variability and sample size? In theory, it would be mathematically possible to combine appreciations of the variabilities of all relevant factors. In practice, no one even attempts this. The usual approach is to pick the factor that appears to have the greatest variability and use it for sample size calcu-

lations, on the assumption that whatever satisfies stringent criteria will *ipso facto* satisfy less rigorous ones. In most textbook tables, the example used for calculating sample size is therefore one that yields maximum variability - usually a dichotomous characteristic (like gender of respondent) where the incidence of each of the two categories is 50% ($p = 0.50$).

Just how far one can go, therefore, in accurately establishing in advance the variability of the characteristics of interest to the evaluation is somewhat debatable. But *some* estimation of this situation must be used in sample size calculations and it is at least worthwhile to be sensitive to the issue.

The other two factors that impinge on the representativity of a sample - that is, its size and the method by which it is drawn - are more directly under the evaluator's control. Within bounds set by available resources, one can increase or decrease the number of units in the sample and change the method for drawing it. Let us consider each of these factors briefly.

Sample size

Generally speaking, in probabilistic sampling strategies, the larger the sample the greater its representativity, since the chances are increased of including appropriate "dosages" of all relevant subgroups. If we were trying to determine the proportion of school dropouts in a population of 100,000 people, a random sample of 100 people would give us an estimation of the percentage of school dropouts that would likely be situated within $\pm 5\%$ of the true value. By increasing the sample size to 500 (and on condition that it be established in the same random fashion), we would have equivalently good chances that the results obtained were within $\pm 2\%$ of the true value.

This relationship is not linear. Further increases in the size of the sample do not bring equivalent improvements in the accuracy of estimate; rather, the improvements tail off. In general, doubling sample size does not halve the margin of error. For similar reasons, the size of the population to which one hopes to generalize results has relatively little impact on the sample size required to obtain a given level of accuracy. If the variability of age in the population remains the same, the sample size necessary to estimate average age within $\pm$ two years is 49 for a popula-

tion of 100 people, 81 for a population five times as large (500 people), 88 for a population of 1000 people ... and only 96 for a population of 50,000 people. In fact, poll-takers in the United States regularly estimate the political opinions of over 100,000,000 voters within a margin of error of $\pm 3\%$ using sample sizes of 2,000 to 3,000 people. The "sampling fraction" (the number in the sample divided by the number in the population) thus has no real significance except in cases where the population is extremely small: see the note on the adjoining tables. Tables 7.1 and 7.2 below give a sense of how sample sizes vary with the size of the population and the level of confidence desired in results (or with the margin of error that the evaluator is willing to tolerate).

In addition, it is important to note that increasing sample size is of little value - at least in statistical terms - if a non-probabilistic sampling method is used. In this case, one cannot be at all confident that accuracy of estimate has improved nor at all sure what the remaining margin of error in fact is.

Sampling method

The third critical factor is the method used in establishing the sample itself. The classic method is random sampling. Like most classics, it is much more often cited or advised than used, for it is quite exacting and cumbersome; but it does define a point of reference in sampling methodology and a way of thinking about the problem. We will consider the "classic" approach first, then briefly review some important variants: stratified random sampling, multi-stage sampling and cluster sampling.

Pure random sampling

Random sampling methods are designed to ensure that every unit in the population has as equal a chance as possible of being chosen for the sample, and that we have a means for estimating how close we have come to that ideal, given certain assumptions. Such are the mathematical conditions for establishing the margin of error in our sample estimates. The following are the five steps in classical random sampling.

1. One first takes care to define the unit of analysis and the reference population for the study.

Table 7.1. Sample size required to estimate a proportion whose true value is near 50%.

		PROBABILITY OF ERROR					
		p = 0.05			p = 0.10		
		Desired level of accuracy			Desired level of accuracy		
		2%	5%	10%	2%	5%	10%
Number of units in sample population	100	(50)*	(50)	49	(50)	(50)	40
	300	(150)	(150)	73	(150)	143	56
	500	(250)	217	81	(250)	176	60
	1000	(500)	277	88	(500)	214	64
	5000	(166)	357	94	1269	258	67
	10000	2000	370	95	1453	265	68
	50000	2381	381	96	1644	271	68

Table 7.2. Sample size required to estimate the average age of a population with an error probability of 0.05 (in 95 out of 100 cases)

		VARIABILITY OF THE CHARACTERISTIC					
		High variability (s = 100) **			Low variability (s = 16) **		
		1 yr	2 yrs	3 yrs	1 yr	2 yrs	3 yrs
Number of units in sample population	100	(50)*	49	30	38	13	6
	300	(150)	72	37	51	15	6
	500	217	81	39	55	15	7
	1000	277	88	41	58	15	7
	5000	357	94	42	61	15	7
	10000	370	95	42	61	15	7
	50000	381	96	43	61	15	7

* In cases where one requires a relatively high level of accuracy with a small population, the formula used to determine the number of units needed in the sample is no longer valid, but a sample of between one-third and one-half of the population drawn at random will generally furnish estimates well within the stated margins of error. Figures derived in this manner are included in parentheses.

**Given a normal distribution, a variance of 100 corresponds approximately to a situation in which 95% of population members are aged from 10 to 50 years. A variance of 16 corresponds to a situation in which the same proportion are between 22 and 38 years of age.

2. Next one must obtain or draw up a complete (and up-to-date) list of all units belonging to the population.
3. The third step is to specify sample size. It will be determined by four factors:
 - a) the number of units that one would ideally like to have in the final sample in order to guarantee the level of accuracy and confidence in results that is desired;
 - b) the number of units necessary to carry out the different types of analysis foreseen;
 - c) likely "mortality" in the sample - that is, the likely proportion of non-respondents or units that, for one reason or another, will be eliminated from the study before its conclusion; and
 - d) the financial and logistic resources available for the evaluation.
4. Next one draws from the overall population a random sample of the number of units dictated by these four criteria. A variety of different techniques can be used for actually creating the sample. These include drawing previously numbered slips from a hat or a box; defining a "sampling interval" n and, beginning with a randomly chosen starting place, choosing every "nth" unit on the population list; and numbering members on the population list and using a table of random numbers to pick those who will be incorporated into the sample.
5. Finally the study itself must be carried out in a way that minimizes any shrinkage or modification of this randomly-established sample, because the further one departs from the initial sample composition, the less the guarantees concerning its accuracy (and the "knowability" of its inaccuracy) will hold. Representativity is always determined in the final analysis by the actual composition of the sample - that is, the characteristics of those who did indeed take part in the study - not by the intended composition or the characteristics of non-respondents! (For this reason, methods for combating or counteracting what is called "mortality" in sample composition - that is, the loss of sample members through non-response or invalid response - take on considerable importance and are discussed further on in this chapter.)

The pure random sampling method obviously has a certain number of practical disadvantages. Three are worth mentioning here. First, it is

sometimes quite difficult to draw up a current list of all members of the population of the study, whether they are individuals of a certain description, groups or localities. Second, if the population of interest is very large or spread over a wide geographical area, the pure random method will often yield a sample that is highly dispersed and therefore difficult to reach with available resources. Randomly sampling all school children in a given country, for example, and then interviewing sample members might require visiting virtually as many schools and locations as there were interviewees.

Finally, one runs the danger with such a mechanical approach of passing over certain types of units which constitute very small proportions of the population but whose experience may have a great deal to say about the program's successes and failures. For these reasons, pure random sampling is rarely feasible in NFE studies. There are other "probabilistic" methods of sampling that are a little more practical and less exacting than the purely random approach, but that still manage to conserve its central characteristic - the guarantee that each unit in the population has an equal chance, or at least a "knowable" chance, of being chosen in the sample. Three will be briefly examined: stratified random sampling, multistage sampling and cluster sampling. Others may be consulted in standard texts on sampling (e.g., Jaeger 1983; Som 1996).

Stratified random sampling

The method of stratified random sampling is essentially based on the relationship between parameter variability and sample size that has already been discussed. If one could identify a series of groups or "strata" within the population that are relatively homogeneous with respect to factors closely associated with the evaluation variables, then the variability of the evaluation variables within these groups should be considerably less than it is within the population as a whole. It should then be possible to compose a population that is representative of each group - and by proportional or weighted addition representative of the entire population - with many fewer units than would have been required by the simple random method.

Let's take an example. Suppose that one wishes to assess the degree of implementation of a public health education curriculum (AIDS

prevention, for example) in adult education centers throughout a given region of a country. Suppose in addition that there are solid reasons for thinking that level of implementation depends particularly on the presence of full-time medical personnel in the locality and on the educational level of the center instructor. If the instructor's previous education can be categorized into two levels - less than high school, and high school diploma or greater - then these two variables define four "strata":

1. centers with high school graduate instructors in localities with full-time medical personnel;
2. centers with high school graduate instructors in localities without full-time medical personnel;
3. centers with non-high school graduate instructors in localities with full-time medical personnel; and
4. centers with non-high school graduate instructors in localities without full-time medical personnel.

These four groups can be displayed in a stratification grid like the one in Table 7.3 and all centers in the region could be classified into one or another of the cells in the table. *Within* each group, the variability of degree of implementation should be considerably less than it is in the overall population, given that all members of the group resemble each other (approximately, at least) with respect to two factors that are closely associated with degree of implementation. One can therefore hope to investigate the implementation of the health curriculum with a sample that is significantly smaller than a purely random one would have been.

Different probabilistic methods may be used for drawing the final sample *within* each of the strata that have been defined in this way. Where it is possible to contact all members of each stratum at reasonable cost, one can randomly select the final respondents. If the groups defined by certain strata are very disparate with respect to some other characteristic - like geographic location - that affects ease of inquiry, then one could perform a second stratification within those groups in order to minimize, for example, the travel required.

As for the number of units to sample within each stratum, two approaches are possible. The first is called "probability proportional to size" (sometimes labeled PPS) and entails selecting within each stratum a number of units equal to the proportion of the related group within the

Table 7.3. Example of stratified sampling frame

		High School Graduate Instructor	
		Yes	No
Full-time medical personnel	Yes	(List of centers)	(List of centers)
	No	(List of centers)	(List of centers)

entire population times the total sample size. In this way, one is assured that the proportions characteristic of the entire population are respected and that results from the different strata of the sample can be simply added together and then extrapolated to those categories of the population as a whole, with whatever margin of error characterizes such estimates. A second approach, with probabilities nonproportional to size, permits the evaluator or researcher to set the subsample size at any level he or she wishes that is consistent with the degree of within-group representativity desired. This approach enables one to study in greater depth groups that may not constitute high proportions of the population as a whole, but may be of particular interest for policy research reasons. They are "representative" of the program in a different sense than the statistical one.

For example, in an evaluation of agricultural extension programs, districts where extension duties have been entirely taken over by village personnel may be distinctly in the minority, yet of great importance because of what they reveal about potential future directions for the program. They might never appear in a proportionately stratified sample, but could be intentionally included in a non-proportional one.

The only *caveat* to remember in this case is that the results obtained from the study cannot be generalized to the entire population without first "re-weighting" the data from the different subgroups in accordance with their actual proportions within the population.

How much reduction in sample size one obtains in this fashion will depend on the quality of the stratification process itself - that is, how closely the stratification factors are indeed associated with the "dependent variable," or the results that we want to assess and analyze (here agricultural extension outcomes). Slonim (1968: 75) cites the case of a study designed to determine the average income of a population of 5000 persons within $\pm 2\%$ where a random sampling method would have required 963 respondents, but a stratified method using seven relatively homogenous groups gave the same level of accuracy with only 170 respondents. In that particular case, the researchers had more exhaustive and accurate information on the stratification parameters than evaluators usually will have; but the example illustrates the sort of reduction in sample size that good stratification can permit.

Stratified random sampling has several advantages over the classic random variety. The restriction in sample size (or equivalently the increase in accuracy obtainable with any given sample size) is one advantage, and the possibility of adopting different final sampling methods within different strata in order to control cost or difficulty of access is another. In addition, with a stratified sampling method, one can explicitly include groups that are proportionally few in number - and thus would in all likelihood not appear in a random sample - but that may have much to offer the evaluation. Suppose, for example, that cases where local religious bodies have recognized the AIDS challenge, though infrequent overall, can be identified and offer particularly interesting models of how the health education curriculum might be implemented. One could include as a third stratification parameter some variable that identifies such progressive religious contexts and thus be ensured of sampling within the innovative group. One would, of course, need to be careful in extrapolating results from the study to "weight" the data in the manner described above.

Stratified random sampling also continues to have some disadvantages inherited from its classical counterpart, however. First comes the necessity of having at the outset a considerable amount of accurate information concerning the population of the study, and notably concern-

ing the potential factors of stratification. Second, a stratified random sample can be virtually as dispersed and difficult to reach as a simple random one, unless a geographical or locational parameter is among the stratification factors. The two remaining probabilistic methods to be discussed here help to alleviate these problems.

Multi-stage sampling

In multistage sampling, one samples by geographical region and/or other relevant dispersion factors *before* drawing the actual study samples to be visited or reached by the evaluation. For example, to obtain a sample of individuals from a population of 50,000 trainees in 2,000 centers spread across an entire nation without incurring exorbitant transportation costs, evaluators may first sample regions within the country - retaining perhaps four out of ten by some reasoned and systematic form of choice related to variability in their characteristics - then sample training localities within those regions, and finally, within the 30 or 40 localities chosen in this manner, sample actual trainees.

One thus obtains a group of subjects that is slightly less representative of the entire trainee population than a purely random sample of the same size would have been, but what is gained from the point of view of economy and feasibility may more than compensate for these losses. Sample definition at the final ground-level stage may be carried out, as noted above, in different manners adapted to the particular circumstances of the localities in question, as long as the method is generally probabilistic and carefully recorded.

Cluster sampling

Cluster sampling is really just a modification of the multi-stage strategy presented above. In the cluster approach, one samples *groups* and then includes in the final sample all members of the groups in question. Thus to draw a sample of trainees, one might sample training centers and then include in the sample all those enrolled. If there were 25 trainees per center on average, one would obtain a sample of 500 trainees by sampling just 20 centers. To draw a sample of centers, one might sample subregions or districts and then include in the sample all centers within the districts in question.

The practical advantages of this approach are obvious, but it does diminish somewhat the representativity of results because, strictly

speaking, the size of the sample is determined by the number of units (i.e., centers) randomly sampled rather than by the sum total of all subunits (i.e., individuals) included within them. Note also that, whereas in stratified sampling one seeks groups that are as homogeneous as possible with respect to the variables under study, here one will seek groups that are as *heterogeneous* as possible, because that is the only way to be sure of including as broad a range of characteristics as possible, given the limitation on the number of groups involved. If you can only interview 15 people for a survey of opinions that may vary greatly by age, you would be well advised to make sure a broad range of ages are represented in the small sample to be established.

Caveats about quantitative sampling

All quantitative sampling procedures that have just been presented need to be interpreted with some flexibility and insight and are seldom sufficient in and of themselves to ensure the optimum "insight yield" of an evaluation. Two reasons for this overall assessment can be cited. The first is that it is scarcely ever possible to apply these methods rigorously and in full conformity with the prescribed procedures. Strictly speaking, as soon as one departs from probabilistic canons by so much as an iota at whatever stage in sampling it may be, the overall methodology used is no longer probabilistic and the margin of error is indeterminate. From that point of view we are all sinners and have fallen short of statistical grace. "Mortality" in sample size (i.e., the shrinkage that results when some intended units do not participate) is a quasi-universal malady. Let it rather be said that the application of probabilistic sampling theories and procedures is a question of degree and of judgment. One generally ends up using, with less than perfect rigor, a set of techniques that is not entirely justified by the situation to which it is being applied, but that nonetheless yields somewhat more representative results than one would have obtained otherwise.

The second point to note is that probabilistic sampling has a built-in tendency to individualize or "atomize" the units in the population that one wishes to study, and to cut them off from each other and from their context. These methods all repose on the postulate that representativeness is a question of addition, or a strategy of "divide and conquer": it is assumed that by summing up the characteristics of a good set of indi-

viduals or units, one can deduce the characteristics of the overall population or collectivity of which they are a part. But truth in evaluation, we have already noted, is as much a question of pattern and relationship as it is of aggregated unit characteristics. These are precisely the respects in which *qualitative* sampling strategies are so important. What one discovers from people in interaction may be quite different, and no less valid, than what one discovers from them in individual interview.

Qualitative sampling strategies

Qualitative sampling methods constitute a second systematic, though non-probabilistic, approach to the sampling problem. If the central thrust of the quantitative approach is to treat all units in the population as equivalent and separate individuals and to ensure that all have an equal chance of being included in the sample, the central thrust of qualitative sampling methods is to recognize that units are not necessarily equivalent with respect to the purposes and needs of the evaluation and that it is important to sample their interactions as much as their unilateral characteristics. Some cases or groups are particularly "rich in information" regarding critical evaluation themes and can offer much more insight than the limited testimony of their demographic characteristics or quantified attributes would suggest. Qualitative sampling strategy is also based on the conviction that a broad range of participants or sites in the program can potentially furnish this sort of rich information, provided that one does not treat them strictly as rapid survey respondents or bundles of measurable demographic attributes, but instead takes the time required to hear and observe what they say and don't say, what they do and don't do, and how they say and do it - and that group formats may be better than individual ones for eliciting such insight.

In short, one could say that qualitative sampling strategy is partly founded on a recognition that, when dealing with human subjects, true representativity is as much a question of how respondents are treated as of who they are. If one is going to invest this extra time and attention in a smaller number of units, however, it follows that one will want to take some care in identifying those most likely to furnish the fullest and richest information. The central theme of the various strategies of qualitative sampling is thus the *identification of information-rich cases*

and the creation of conditions under which those resources can be tapped and their yield in insight cultivated. Four related techniques merit particular mention: intentional or purposive sampling, snowball sampling, elite sampling, and group sampling.

Intentional or purposive sampling

The most common approach in qualitative sampling can be summed up with the adjective "purposive" or "intentional." In an intentional sampling procedure, evaluators and their informants identify categories of units or individual cases or localities where key dimensions of the program in question can be experienced, studied or assessed in particularly rich detail. Centers where trainees have had the opportunity to manage and plan courses themselves might be one example, or places where local authorities have been particularly active in support of the program (or perhaps particularly hostile to it).

If one is to identify such "information-rich cases," a concurrent criterion in purposive sampling must obviously be the possibility of access to this depth of experience, or what qualitative researchers call the issue of "entry" (a topic more thoroughly explored in Chapter Eight on qualitative evaluation). There is no use selecting highly interesting targets for inquiry if, in fact, the particularly informative aspects of their experience are impossible to observe or record and one can only get concerning them the minimal indicators typically collected by a quantitative evaluation. For example, it might be very revealing for us as evaluators to interview the president or prime minister of the country regarding educational and social policies impinging on the program we are evaluating, but the likelihood of actually being able to conduct such an interview is very slight. We don't enjoy the requisite "access" or "entry."

Snowball sampling

Snowball sampling is simply a variant on the purposive approach where evaluators - in a manner typical of qualitative research strategies - use their experience in the field and their interviews with a first tier of respondents to generate new units for inclusion in the sample. This method, which obviously bears a relation to the investigative journalism

technique of following leads, is particularly useful and appropriate in those numerous circumstances where the evaluator does not know exactly who would be the best informants or which cases would provide the richest information, but where she or he suspects that key stakeholders would in fact know this full well. One then takes care to choose as accurately as possible a first set of units to investigate or a first group of subjects to contact and includes in the evaluation method itself ample opportunity to query respondents about who or what else might provide rich insight into the operation of the program being evaluated. Each person or locality may thus lead to one or more others. In fact, the leads tend usually to multiply at such a rate that the researcher is faced with choices - what qualitative researchers call decisions about "inclusion and exclusion." And these choices are an important tool in shaping the inquiry.

Sample size in this sort of scenario is determined less by statistical probabilities than by one's sense of how much has been discovered. Qualitative researchers speak of the notion of "saturation." One ideally continues this sort of inquiry until (or just beyond) the point where the information obtained *begins to become systematically repetitive* - the point of "redundancy," in qualitative jargon. This provides an indication that the researcher has begun to gather all the insight and experience from the field that he or she can with their present conceptual tools.

Elite sampling

Another variant on the same approach involves explicitly identifying those people who seem likely to have the most varied experience with the aspect of the program under consideration or the greatest insight concerning it - whether they are among the program's intended beneficiaries, its present and former staff, its directors, or outside stakeholder groups - and then interviewing them in depth, usually over an extended period of time.

Of course, the word "elite" has a distinctly different meaning in this connection than it does, say, in sociological or political parlance. It obviously does not necessarily refer to interviewees from the upper class, but simply to those "favored" by particularly deep and insight-laden contact with the program, or who are in a position to put it in interesting historical and comparative perspective. Individuals who were particu-

larly instrumental in the creation of the program may fall in this category, as might those who have many years of experience with it or possess special knowledge (technical, financial, political) about poorly understood dimensions of it. In short, they are in some sense people who have already been gathering and processing valuable "data" on their own - the data of personal experience - and who therefore offer evaluators a shortcut to key understandings. Gaining "entry" to such key informants is another question, and one to which we shall return in the chapters on qualitative evaluation methods, for one is not likely to identify these resource persons by simply drawing straws from a hat or picking from a table of random numbers, and actually "tapping" their wisdom requires more than writing their name on a list.

Group sampling

Yet another variant involves sampling *groups* of people that will be interviewed or engaged collectively rather than individually, or pulling together respondents from different locations - in the manner of a panel or focus group - who will then be involved together in providing and analyzing information about given aspects of the program. The dynamics of group sampling and group interviewing are obviously quite different from those of the individual variety, but this technique can provide insights that the more "atomized" approaches of classical surveying simply can not.

In all of these variants of qualitative sampling methodology, the concern is to apprehend *pattern* as much as individual characteristics and to view aspects of the program within their context. In addition, one may be looking for the leading potentials of the program as much as its current characteristics: i.e., an indication of where it might go, how it might best fulfill its objectives or mission. Such insights cannot be determined on a purely statistical basis but are no less important for that. "Vision" is a critical category in most current discussions of organizational and corporate renewal, because it is often the factor that motivates organizational members to excel and that pulls together otherwise disparate sets of people in effective joint effort.

Yet the best vision is not one imposed from outside, but one found within. Such vision is always more of a nascent phenomenon than one evident on every street corner, something latent here and there in the

program, implicit in little-noticed innovations and practices, but half-hidden until circumstances or leadership reveal it. A sampling strategy that would uncover it cannot be purely probabilistic, because renewal and vision are statistically improbable - but happily not infrequent - events. They must be ferreted out. Sampling experts take heed: Good programs are involved in some sense in making the improbable probable, and discovering how requires varied means of selection.

Qualitative approaches to generalizability of data

Qualitative studies and qualitative sampling methods are frequently taxed with producing "interesting but non-generalizable" results, to such an extent, in fact, that graduate students proposing qualitative research are often careful to say at the outset of their prospectuses that they will make no claims of generalizability for their findings. The perspectives proposed in this chapter on sampling and those to follow in Chapters Eight and Nine on qualitative evaluation technique should throw some doubt on the issue, however. For one thing, some of the most important characteristics of a situation or of a given population may be nearly the same for all components of it (low variability and hence low requisite sample size) and yet so well hidden from sight or consciousness that only, in-depth qualitative approaches enable the evaluator to apprehend them. For another, through deriving estimates of the prevalence of given individual characteristics in a population is potentially the strong suit of quantitative sampling, evaluators are finally more concerned with patterns and meaning than with head counts *per se*, and to grasp these requires identifying, experiencing and/or creating situations where people construct such patterns. That is a task of qualitative selection.

What are we to make of the issues of representativity, margin of error and sampling bias in the qualitative domain? Once again, qualitative sampling takes a different approach to these issues than does the quantitative researcher. The qualitative researcher asks what we *mean* when we say that data are representative. Representative of what?

Quantitative researchers, would respond that such data constitute a relatively accurate or faithful image of the population of interest and that we can say with some confidence just how accurate they are. The qualitative researcher, on the other hand, would maintain that representativity is a more complex question than these remarks imply. People and

programs have surface characteristics and they have deep characteristics. What finally interests us in evaluation are *patterns of factors and characteristics* that help explain why program participants and stakeholders do what they do and how results are or are not achieved. Those patterns arise from deeper levels of meaning and motivation in people's individual and social lives. They do manifest themselves in surface phenomena, but are perhaps best discernible in the details of interaction and reflective expression, rather than in the gross outline of performances and demographic characteristics or in the completion of responses to closed-ended surveys. If we want our results to be representative of this deeper reality - to offer a faithful image of what is really driving the program - then we have to sample and inquire in a fashion that allows us to get beneath the surface and plumb some of these depths, in a way that enables us to inspect the detail of experience with a care that will reveal its fine (as opposed to its gross or transitory) structure. In the final analysis, the point is to get representative *information* - i.e., information representative of what really interests us - not just representative units.

Mixed methods

It should be clear from the foregoing discussion that I consider quantitative and qualitative approaches to sampling much more as complements than as competitors. Notes of competition creep into this dialogue largely because quantitative - or "probabilistic" - sampling has heretofore been considered the *only* valid kind. No one likes to share absolute power, and those who are excluded from or victimized by it generally resent the fact. But it seems clear that both approaches are needed and that both approaches have in fact always been used. The quantitative probabilistic methodology does provide the best handle on representativity of units (given certain assumptions), but not on patterns and meanings, which are finally the factors of greatest importance to evaluation. The qualitative information-rich methodology gives greater depth and therefore the best "representation" of underlying issues, but it is subject to bias and partiality which, finally, only a probabilistic cross-validation can check.

As a result, it will often be useful and is entirely feasible to combine the two approaches. There are multiple ways of doing so. In general terms, one can start with a quantitative methodology and then incorpo-

rate a qualitative one - as when one randomly (or by stratification) samples sites and then seeks interviewees within them by elite or snowballing techniques. One could equally well pursue matters the other way around - as when an evaluator chooses sites on intentional criteria and then samples probabilistically within them. The underlying criterion is to maximize insight from as representative a group as possible, or, more accurately, to obtain *information* that is as representative as possible of the true situation of the program and its leading potentials.

Other important dimensions of sampling

We have spent most of the space in this chapter discussing demographic and geographical sampling techniques - both quantitative and qualitative ones - because they have been the most developed and emphasized in the research and evaluation literature. The same distinctions made above throw light, however, on the other dimensions of sampling mentioned at the outset of this chapter.

Temporal sampling

First, consider the issue of "temporal sampling," which is seldom given the attention it deserves. Programs and their participants are highly subject to complex time rhythms. Some rhythms are built into the nature of the program, some derive from events and patterns in the social context and physical environment. (Think of the number of rural training or literacy programs - not to say primary schools - that have functioned poorly because they took inadequate account of agricultural seasons.) Such rhythms inevitably modify the way in which the program operates and so affect the data that may be gathered about its operations. Explicit account needs to be taken of these factors in evaluation design.

There is therefore every reason to define a "sample" of times and junctures at which program results will be assessed and program processes observed. Both quantitative and qualitative sampling criteria apply here. On the quantitative side, though budgetary resources generally limit evaluators to small time samples, one will nonetheless be concerned in the multi-year calendar of monitoring and evaluation activities to be sure to experience and collect information on the many different phases

through which the program evolves in each locality. On the qualitative side, evaluators should be concerned to identify those "high points," important events and critical passages where program dynamics are particularly well exposed and to pay particular attention to these. The observations here about temporal sampling also should make it clear why longitudinal studies and evaluation *programs* that continue over time are more likely to obtain truly representative data than are one-shot assessments. Only time allows us to establish a better sample of the universe of events that determine the success or failure of programs. Though single studies seldom allow such luxury, a career of evaluation work does.

Thematic sampling

The thematic dimension of sampling is always a prime subject of debate, though it is rarely labeled a sampling problem. It is rather discussed under the guise of choice about which particular aspects of a program will be evaluated or studied. The "frame" for this dimension of sampling is really given by the program model and by current understandings (or misunderstandings) of the context in which it is operating. Which of these many dimensions of the program will receive most scrutiny?

Choices here are strongly influenced by political and administrative necessity, but evaluators may nonetheless have, over time, a good deal of latitude in choosing where they will direct their attention. The criteria at this level are probably mostly qualitative and strategic: what are the current bottlenecks in program operations, what the dimensions of results that seem to promise the most for the future?

One obvious criterion in thematic sampling is the level of interest of different issues for decision-makers and important stakeholders and thus the potential of these topics for promoting constructive discussion about the program's future. In any case, in this area, as in the realm of quantitative sampling, one generally does best to start with a broad "census of population units" - i.e., an inventory of potential themes of inquiry - and then to select from among them those that best meet key criteria. Written evidence of this process, and of the participation of stakeholder representatives in it, is often very useful later to justify the choices made in evaluation design. It is one thing to be criticized for methods used in an

evaluation, and quite another to be taxed with having addressed the wrong questions to begin with!

Theoretical sampling

Theoretical sampling, which concerns choice of conceptual framework and disciplinary perspective, is generally given less prominence in evaluation than in research but is worthy of more attention than it receives, if only because theoretical perspectives are implicit in all evaluation methods. The basic question to ask is the one posed by researchers in what are commonly called the "foundations of education" (educational psychology, educational linguistics, the sociology and political science of education, the economics and anthropology of education, and the history and philosophy of education): what different

Box 7.1. Theoretical sampling: Lessons from Machiavelli's The Prince

Niccolo Machiavelli, the medieval Italian political thinker and author of *The Prince*, wrote in his work that a sovereign benefits most from having advisors who contradict each other, because he or she can then choose among these divergent viewpoints the alternatives that best meet the needs of the crown (Machiavelli 1996). There is a lesson here for evaluators and for those who sponsor them. While it may at times be in the interests of outside funding agencies to conduct unequivocal (or "univocal"!) evaluations that confirm the decisions they wish to make, evaluators know that multiple sets of data could be gathered about any given program, and that any given set of data could be interpreted in multiple ways; and program directors generally prefer to keep their options open in deciding what lessons to draw from evaluation and how best to promote program performance.

One manner of guaranteeing this latitude is explicitly to include within the set of evaluators people with different theoretical and ideological orientations and nurture thereby a genuine debate about what the program has accomplished and where it should be going. This tactic of "theoretical sampling" is related to an evaluation strategy first popularized in the United States in the early 1970s - "adversarial evaluation," which modeled the conduct of evaluations and the presentation of data on the example of legal cases, where opposing points of view are compared and there is always a prosecuting and a defense attorney! (Wolf 1975; Levine 1978)

"prisms" or interpretive frameworks might be used in analyzing program operations and in understanding their dynamics?

Each of the disciplines mentioned offers perspectives that may be of great use in evaluation, and many existing evaluation procedures are directly or indirectly derived from the ones named. Most evaluations are based, in fact, on a blend of several perspectives, which is one reason why it is pertinent to talk of theoretical sampling. It is increasingly common to assemble "cross-disciplinary" teams - composed of economists, sociologists, educational psychologists and others - to perform large evaluations. (The tactic may not be as new as it seems: see Box 7.1!) The importance of progressively refining and strengthening these underpinnings of practice is one of the best arguments for the "liberal education" of evaluators, and one to which we shall return in the concluding chapter of this book. The better one understands different ways of investigating truth, the better one can interpret the data at hand.

In any case, the issue of theoretical sampling and the blending of multiple interpretive frameworks has a double importance for educational evaluations. It serves to sensitize us to the necessity of adopting and developing a composite analytical perspective in evaluation (or of recognizing and refining the one that we already have); and it lays a basis for the creation of inter-disciplinary or inter-agency evaluation teams. To see in depth one needs stereoscopic vision, and that cannot be gotten with less than two distinct perspectives!

Conclusion: The universality of sampling concerns

Mixed methods and triangulation are central themes for evaluators. The prime concern in evaluation design is how to put together the mix or combination of methodologies and studies that will yield the fullest and most useable insight regarding the program. The first step is to widen the possibilities, a concern throughout the different chapters of this book. The second is to pick among these options the ones best adapted to given evaluation exercises. And that is, once again, a sampling issue. Generally speaking, the broader the options, the better the choice eventually made - and the more divergent the approaches that one can reconcile within a coherent design, the stronger will be the results of the evaluation.

The spirit of sampling penetrates every aspect of research and evaluation because *making choices among alternatives* is the task at every step along the way. We saw in Chapter Five that concerns about reliability and validity fundamentally involve sampling issues. So in fact do most aspects of evaluation design. We are always doing less than we might have wished or would have hoped, though frequently more and better than has been done before. In these circumstances, balanced choice is critical, and that means awareness of alternatives and their variability, clear criteria, a systematic procedure for selection, and careful follow-through. Sampling methodologies deal with just those issues.

Notes

¹ Fuller detail on the techniques discussed here may be found in texts like Jaeger 1983; Krishnaiah & Rao 1988; and Som 1996.

² Had they wished, of course, to establish within ± 10 centimeters the average height of the world's adult population, they would have needed a much larger sample drawn in random or carefully stratified fashion - because in that case the *variability* of the factor "height" would have been much greater. Its values would no longer be "yes" (> 50 m.) and "no" (< 50 m.), but instead a very long string of decimal-place figures ranging from about 1 m to about 2 m 20.

References

- Jaeger, R.M. 1983. *Sampling in Education and the Social Sciences*. New York: Longman.
- Krishnaiah, P.R. and C.R. Rao. 1988. *Sampling*. Amsterdam, NY: North Holland.
- Morris, S. 1968. *Sampling*. New York: Simon and Schuster.
- Slonim, M.J. 1966. *Sampling: A Quick, Reliable Guide to Practical Statistics*. New York: Simon and Schuster.
- Som, R.K. 1996. *Practical Sampling Techniques*. 2nd ed. New York: M. Dekker.

CHAPTER 8

QUALITATIVE EVALUATION METHODOLOGY: A DESIGN ON MEANING

If the underlying goal of evaluation is to maximize useful insight into program operations and impacts - and to optimize the use to which "insightful" data are put - then quantitative information alone obviously will not fill the bill. Numbers by themselves just don't aggregate into insight or decisions. They must be fitted into some kind of framework that makes them meaningful, that interprets their direction, and that interpolates between them or extrapolates around them some of the contextual detail about program operations and effects on which good decision-making depends.

A statistical table or graph constitutes one step in that direction. Numbers there gain added meaning from the columns and rows in which they are placed, or from the ways in which they are juxtaposed, arranged and portrayed. We often forget that the significance of a statistical table comes as much from the way in which the data are arranged as it does from the numbers themselves. For similar reasons, the literature on indicator systems (e.g., Chapman, 1992; Easton et al, 1990) generally distinguishes three levels of quantitative information: (a) *raw data*, which are the simple results of counting or measuring; (b) *statistics*, which are uses and aggregations of these numbers that summarize characteristics of broad groups; and (c) *indicators*, which are selected or compiled statistics considered to be expressive of some key attribute of a program or a program-relevant situation. Moving from one level to another of this hierarchy, as researchers and evaluators repeatedly do, involves considerable processing and framing of the original data.

In the first chapter of this section, on "Data Quality," we discussed the epistemological issues that underlie evaluation efforts and characterized four different perspectives or paradigms that have differing implications for the way in which evaluation work is conceived, designed and conducted. The challenge of alternative perspectives to the positivist and post-positivist approaches that have remained dominant in official circles comes down to pointing out that *theory generally precedes data collection; it does simply not follow or flow from it*. Theoretical perspectives

are already implicit in decisions about how to define data, which data to collect and how to collect them. The humorous example in Box 8.1 makes the point particularly clear.

Who defines the data?

The meaning of quantitative data depends on the way in which the data are framed. But there are many different ways in which programs can be understood and interpreted, and thus a number of radically different choices that could be made about what to treat as data and how to count, aggregate and display this information. Moreover, each stakeholder group may have its own understanding of the program and the issues affecting it. Whose framework and interpretation will be used? Who gets to define the data? Whose information will count as information? How will the principal interpretive framework be chosen or devised?

These questions are especially critical in program implementation and evaluation, because one of the prime objectives of most social programs is to ensure successful dissemination and adoption of the innovations that are being introduced - that is, to promote genuine "ownership" of the new strategies by program beneficiaries and thus increase the likelihood that improved approaches will continue to be

*Box 8.1. What do you count?
Or how assumptions shape numbers*

Kevin Harris (1980: 122) tells a hypothetical tale that shows how theoretical perspectives are embedded in numbers. Suppose, he says, that an "objective" observer is charged with counting the human population present on the local football field at exactly 2:00 PM of a given day. This seems like a relatively simple and straightforward task. But if there are three pregnant women present, one person who is sleeping half in and half off the field, and another who has unfortunately just had a stroke and passed away, the latent problems in this simple counting assignment become apparent. The results will depend entirely on how the concept of "person" or "people" is defined (explicitly or implicitly) ahead of time. Is a pregnant woman counted as one human being or two? How much of a person's physical being must repose on the field for him or her to be included? Is a recently deceased individual to be considered part of the population for these purposes or not?

applied once the program itself has come to an end. Ownership supposes, however, that program beneficiaries have the opportunity to modify the proposed strategy, adapt it to their particular circumstances, and come up with hybrids of the innovation and their existing patterns of behavior that have meaning for them and enhanced survival value in their environment. The results of any intervention of this sort therefore have as much to do with the way in which the program theory or model is modified and assimilated by the host culture as they do with simple output measures. In a sense, the initial strategy or program theory of an educational intervention can be seen as a catalyst, a stimulus which should lead all involved to develop locally adapted versions of the methodology. These "local versions" may have a much better chance of being adopted and durably applied by the target population than the one initially introduced from the outside.

If this is the case, *sensitivity to alternate ways of understanding program objectives and framing data about program performance* is obviously critical to accomplishing the "process" goals of the intervention. In fact, programs are partly designed for the precise purpose of eliciting and articulating these differences and of seeking a dynamic middle ground. A method of evaluation that glossed over or submerged differences in perspective would therefore do little service to the long-range objectives of the program and would risk providing little insight into the factors that explain why observed results have been achieved.

Yet the dependence of quantitative data on the way in which they are framed is rarely either recognized or taken into account. Quantitative data are such a part of official "culture" in most agencies, businesses and major social institutions that questions are rarely asked about their validity. Whatever the ultimate validity and reliability of numerical data, quantification of information is a sort of widely respected ritual with its own norms and procedures. John Meyer (1972) once pointed out, for example, that the transnational culture of "modernity" which has been spreading worldwide over the last half of the 20th century virtually requires of self-respecting organizations and institutions that they employ economists on their staff or engage in varieties of economic analysis as part and parcel of their planning, even if those analyses have little evident relation with their institutional missions and irrespective of the fact that no one within the organization is able to interpret or use them. Economics has become a widely-accepted (and expected) language for strategic planning, and planning documents that fail to use it

risk not being considered valid or genuine by the boards and directors to whom they are addressed. In a similar vein, Campen (1986) points out that statistical analyses became such an obligatory part of legislative presentations before the United States Congress in the 1960s and 1970s that policy proposals bereft of quantitative argumentation had little hope of passage. Quantification essentially drowned out any other form of evidence or argument.¹ Evaluators would do well to keep in mind the cultural and political meanings of quantitative information.

Opening up the black box

Qualitative approaches to evaluation and research can provide opportunities to investigate the meanings attached to events and to "data," the patterns that people attribute to them, and the way things "fit together" in the minds of stakeholders. In fact, the results of program evaluations carried out in the 1960s and 1970s have had much to do with the tardy but increasing acceptance of qualitative methods. Time and again these evaluations demonstrated that large-scale social programs had not achieved the hoped-for results or had produced effects different from and those anticipated. The question each time was "why?" Numbers alone could not provide the answer; attention focused increasingly on the process of implementation itself. Clearing up the mystery and increasing program effectiveness required, it was decided, a better understanding of the way in which programs were actually carried out and of the "mutual adjustments" or conflicts that intervened between their official promulgation at a national level and their real delivery in the field.² The problem with quantitative assessment procedures which focused on inputs and results was that they treated the vital processes of program implementation rather like a "black box" and so provided little insight into how to remedy the problems that they highlighted. Program evaluation has therefore not just been a beneficiary of the growing attention given qualitative methods; it was historically one of the causes of this phenomenon.

I want in this chapter and the following one to make at least a brief excursion through qualitative evaluation methodology and the role it can play - both in isolation and as a complement to quantitative approaches - in increasing the "insight yield" of evaluation. If not further specified, qualitative methodology has an unfortunate tendency to become a sort of

residual category: a rubric and rallying point for everything different from quantitative research and for anyone who feels uncomfortable with numbers. Such a vague and sweeping approach doesn't do justice to the sophistication and artistry of qualitative evaluation. This chapter evokes some of the main attributes of qualitative evaluation design; the following one is devoted to presenting selected aspects of its methodology.

Thick description

What are qualitative data? The category is usually defined by three types of information: interviews, participant observation, and content analysis of documents (or, more generally, "artifacts"), as well as by various combinations of these three categories. From the point of view of their uses rather than their constituent parts, however, qualitative data might be defined by two features which deserve fuller comment: thick description and meaning patterns.

"Thick description" is an ethnographic term for information on phenomena and events that includes not only richly detailed perceptions of their characteristics but also a good deal about their context and history, about how the people most affected experience and view them, about what other things seem most closely associated with them. A "thick description" involves much more varied information about an event or person or object than we normally perceive in the course of everyday encounters and observations, and certainly more than is indicated by numbers or typically presented in program reports.

It seems fair to say that we usually miss, screen out or take for granted a good deal of what is actually going on around us (some psychologists would say as much as 80%), even in environments with which we are familiar, let alone in those that are more foreign to us. For one thing, our sensory organs can only handle and unscramble a certain number of stimuli at one time. As a consequence, perception is in good part a task of *selective attention*, and therefore of selective suppression of what is perceived to be excess information. For another, we may not want to "hear" or "see" certain things that would be disruptive of our values or our sense of ourselves.

Thirdly, much of what is happening around us is expressed in languages that we are not even able to interpret or decrypt. The example of cross-cultural interaction obviously illustrates this point, but the same

idea applies equally well much closer to home. Even within a given culture, men and women, adults and children, laborers and managers may not have the same language; and much of what members of one group are saying - in their words, actions, or "artifacts" - may be unintelligible to members of another group. Think also of the example of a person raised in the countryside who hears and sees whole realms of meaning and types of information in the natural environment that a typical city-dweller misses, or one bred in the city who understands from its sounds and sights clues to what is happening that a recent rural immigrant cannot decipher. A West African proverb reminds us that "the eyes of the guest may be as wide as dinner plates, but he still does not see the town as it really is."

"Thick description" implies the record of a perception that is much broader, deeper and more detailed than what we normally produce and that captures many of these implicit meanings. The importance given to it in qualitative inquiry is based on the notion that the answers to many of our questions are already "out there" in the environment, if we could only open our eyes and ears wide enough to perceive and examine some of the data of experience that we have been screening out. Yet very simple elements of this sort may do much to enrich an evaluation. (One illustration of these simpler uses of qualitative description is found in Box 8.2.)

Historical inquiry

One variety of "thick description" that immediately gives greater depth to an evaluation consists quite simply of historical inquiry into the origins and development of programs and the chronicle of events that have happened in and around them or among the stakeholder groups involved. History is a critical dimension of the present that we too often ignore and that is frequently given short shrift in administrative reports or largely quantitative evaluations. And yet not only does it explain much about the present, but the act of explicitly taking people's lived experience into account gives us greater sympathy for them and gives them greater confidence in the evaluation itself.³ Much that may seem abnormal or deficient in a program appears in an entirely new light once its history and the forces at play are understood. As the Sahelian proverb puts it, "Whatever drove the mouse to jump into the fire must have been hotter than fire itself!"

*Box 8.2. Thick description:
An example from a Malian literacy evaluation*

The term "thick description" is often reserved for full-scale ethnographic accounts of observation or in-depth interviewing of key informants, but elements of it can usefully be employed to give greater depth and more of a corrective sense of reality within evaluation reports. The following paragraph is excerpted from an evaluation of the functional literacy campaign in Mali and was based on careful local observation in a number of localities (République du Mali 1978: 136, translated from the French). Note that quantitative information is used, but it is placed in a descriptive context.

"It is important to picture the conditions in which teaching and learning are carried out. The instructor generally stands before a dozen students out of the thirty supposedly signed up for the courses. Six or eight of these are his regulars: those that attend 80% of the class sessions; but several won't have been in the last class and almost all of them will be absent for a week or more at a time carrying on dry season commerce in other villages or towns. The students may range in age from 15 to 50 and have very different levels of accomplishment in literacy. Some of them aren't yet able to decipher individual words, while a few others are at a level not much below that of the instructor. Four or five are former primary school students who left classes at different points short of the sixth year.

Lighting is typically weak: several storm lanterns burn in the front third of the mud shelter. The older students and those afflicted with river blindness - perhaps a third of the group - have great difficulty making out what is written on the faded blackboard, which is perched precariously on a broken easel in the front corner. Yet the instructor has thought to break his student body into groups, with one of the most accomplished students assigned to each, and these assistants listen and correct while the others attempt to decipher articles from last year's book of readings. No new material has arrived yet this year and the instructor says that the center has not succeeded in subscribing to the national Bambara-language newspaper."

For related reasons, participatory evaluations often begin with a request for program beneficiaries or other stakeholder groups to recount in their own words how the program developed, who initiated it and/or most influenced its implementation, what happened from year to year, and why. Just listening to this kind of chronicle - and particularly listening in an active and attentive way - is a research act of great significance. Piecing together the recitals of different stakeholder groups

into a synthetic (or perhaps contrastive) "life history" of the program quite literally puts all phases of evaluation that will follow into a much richer and deeper context.

Meaning patterns

Meaning patterns are the second key feature of qualitative data, and our understanding of them is typically built up from thick description. Meaning patterns are the configurations that determine how we interpret and categorize our experience. Educational psychologists have long remarked that people have much greater ease learning and recalling information if they use or are given "mental schemata" for organizing it—that is, structures that help us to arrange information into meaningful wholes or to relate it to what we already know. Impressions and "data" are not individually stored in and retrieved from our brains; rather, they are *configured in patterns*. Related items are linked together by temporal or logical association. Smells may evoke for us strings of memories; images are connected with strong feelings and values; events are associated with their perceived consequences; and we remember most easily dates or facts or numbers that are related to things we already understand or value.⁴

The point is simply that the contents of our memory and of our consciousness are structured in complex functional ways, though ones that we continually reshape to some degree on the basis of experience. Equally important, these structures are also *the medium through which we perceive new experience*. It was noted above that our sensory organs can only handle a fraction of the stimuli and perceptions that they receive from the environment and that we must therefore both select and suppress information on an ongoing basis. Existing patterns of knowledge in our minds naturally become one prime means for making this selection, if only because it is much easier to "grasp," store and subsequently retrieve information that fits into existing mental configurations than it is to revise these or add new ones. Incoming impressions are filtered through and compared to the patterns of meaning and association already layered into our consciousness. We recognize those (like sounds in a familiar language or dialect) that fit with our own patterns and, unless a conscious effort is made to the contrary, largely disregard those (like

"gibberish" in unknown foreign languages) that cannot be related to anything we know.

Cultural and social "algorithms"

Meaning patterns are thus the constituent elements of our whole socially and culturally-determined outlook on the world and might be (rather woodenly) represented by a set of algorithms: conditions *a*, *b* and *c* exist because of historical factors *x*, *y* and *z*; if you act in manner *m* in the situation *s*, the result will be either *q* or *r*; people *p* typically treat us in manner *t* because their historically- or divinely-determined nature is *n*; and so forth. Obviously the real "algorithms" by which we live are a good deal more complex and flexible than this and may vary over time and among members of the same society, but our world is truly built of such meaning patterns. They explain what is significant to us, on what it depends, what kinds of activities will be potentially effective in achieving our ends, and how constraints encountered in everyday living may be dealt with. And yet these patterns are generally invisible to us.

An example or two will make the point clearer.⁵ A classroom is a space occupied by people who meet, work, play and interact with each other. But they generally interact in very set and prescribed ways (how set and prescribed depending in part on the type of educational institution involved) - ways which the participants themselves probably take for granted or of which they are not consciously aware. Normally at the start of any given class in a traditional university setting, for example, students and professor walk in and sit down or stand in accustomed places as if this were the most natural and unproblematic thing in the world. But if I, as professor, walk in and sit in the midst of a group of students in the back row, people immediately sense that something is wrong. Some will be amused, others irritated, still others pleased - but it is evident to all that one of the "norms" or rules governing this particular social situation has been breached. It is these invisible rules determining how people define and experience situations and the roles they play in them that are of particular interest to a qualitative researcher, for they determine much about what subsequently happens.

Roles and rules of behavior at baptisms, in market interactions, or in funerals and situations of bereavement, are for the most part equally structured and "invisible." In fact, "classroom," "baptism," and "funeral" are three in a long repertory of events and settings that members of any

culture learn to handle during their apprenticeship and socialization into adult life. In most cultures, someone playing loud rock music at a funeral would immediately be seen as violating a norm, as being "out of place." Yet in the Creole culture of the Caribbean and Louisiana, jazz performance is a regular part of funerals.

Meaning patterns often therefore come into consciousness only when they are violated or put into question, which is why cross-cultural encounter is one of the principal experiences making people aware of the norms and rules that govern their behavior and environment (but that do not govern the behavior of those "others"!). Bargaining in commercial transactions offers a good example. Some cultures define it as normal, even expected; others rule it out in a wide range of situations. An individual from culture A who tries to bargain down the price of merchandise in a store governed by culture B soon feels that he or she is dealing with an unfamiliar set of rules - as do the store personnel.

Meaning patterns are constituted in good part of just such largely unconscious ways of defining situations, such norms of behavior and configurations for the furniture of life. People are constantly constructing and rebuilding these scripts in their everyday existence, repeatedly expressing them through speech and activity, and transmitting them - explicitly or implicitly - to the young and to other cultural newcomers. Though many of the fundamental patterns are passed down from generation to generation, we are not speaking exclusively here about deep-seated cultural rituals - and even these must be "operationalized" in more localized and mundane social applications if they are to survive. In fact, every workplace, and every training center or school has its own culture built up from a mix of broader social norms, constraints and requirements peculiar to the activity at hand, and the personal idiosyncrasies and life histories of those involved. Discovering this culture and describing how it functions is a central preoccupation for the qualitative researcher or evaluator. One sees readily why anthropologists, with their close attention to cultural traits and differences, have typically been in the forefront of work on qualitative research. Quick perusal of the literature reveals that the tradition of qualitative evaluation in education has grown largely out of ethnographic study of classrooms and other educational settings - as anthropologists have moved from studying solely "exotic" cultures (exotic to them) to analyzing the many intricacies of their own.

Evaluating meaning patterns

What determines the value or adequacy of a meaning pattern? We might cite three characteristics. First would come its *motivational potential* for the community or set of stakeholder groups involved: to what degree does it stimulate their imagination and elicit their commitment? Second might be its "*productive potential*" - its ability to make use of available resources and technologies and draw optimum yield from them. A third and last characteristic would be its "capaciousness" or *absorptive capacity* - its aptitude to take account, and to render account, of the full range of group member experience. Most Europeans of the 15th century shared a pattern for interpreting and understanding geography that was comfortable, motivating in some respects (like its relation to contemporary religious doctrine), and accommodating of everyday land-bound experience: they believed the world was flat. The view proved, however, poorly adapted to the navigating challenges of the age of exploration and little able to take account of the experience of Columbus, Magellan and the host of adventurers on the forefront of European imperial and commercial expansion.⁶

An important distinction should be made between (a) the breadth of experience and facts to which a group has access, and (b) the adequacy of the meaning patterns built on this experience. Evaluators should be open to the possibility that people with less access to experience and data may in fact have built patterns of meaning better adapted to facing the future or dealing with certain aspects of it than those with greater information resources. Cheikh Amadou Kane has maintained, for example, that the Wolof language of Senegal is better adapted to the presentation and understanding of Einstein's theory of relativity than most European languages by virtue of its internal logic and structure. Local participants in programs may have in some circumstances a better "fix" on what is right and wrong with intervention methods than do national staff, despite their lack of system-wide information. The criterion of worth in evaluation is depth of insight, not quantity of data. The necessity of openness to such possibilities makes clear the importance of taking meaning patterns seriously and investigating how those that are implicit in a given program have interacted with the prior frameworks of its intended beneficiaries or prime stakeholders. It is at this level that a program proves its "sustainability."

Programs as meaning patterns

For in fact programs themselves are essentially shared meaning patterns. There is no physical entity out there called a "program" that we can go to and weigh or measure or even interview. There may be local offices that devote much of their effort to activities funded by the program or labeled with its name, as well as personnel who have been assigned to it. But the program itself is a shared pattern of meaning and commitment that we create through our intent (the "program theory") and our action. It is a mental construct designed to motivate particular sets of people and to direct their energy in particular ways. Once that common pattern is gone, there is no more program, though buildings and personnel may continue to exist. Apprehending how people understand program goals and processes is therefore the heart of evaluative inquiry. Learning to read the patterns of meaning and causal understanding, and to discern how well the patterns proposed by the program have been grafted onto - or crossbred with - existing ones, are key competencies that qualitative evaluation both requires and helps to develop.

Discerning meaning patterns is, however, no mean feat. Our understanding of them is built out of and depends upon thick description. Since new meaning patterns are in many respects foreign to us, we are not likely automatically to perceive them. That is not the way in which we would organize the data of experience. To pick up these other data and the variant ways in which they are organized, we need a much more ample perception and description of the environment and of different parties' "constructions" of it than we normally bother to develop. It is the richness of detail about what events and phenomena are like, how they occur, which of them are associated with which others and how stakeholders view and relate them that provides a basis for discerning meaning patterns and for recognizing cause-effect relationships and values that are markedly different from our own or from what we normally perceive.

The meaning patterns that we reconstruct through thick description constitute "theories" about how some domain of social or personal life operates: who does what, why and with what effects. Qualitative researchers (beginning with Glaser & Strauss 1967) tend to call this procedure "grounded theory" to indicate that it is one in which constructs and patterns are inductively derived from very plentiful exposure to the details of daily life, rather than deductively drawn from existing bodies

of research or theory. The word "theory" has a somewhat different meaning in Glaser and Strauss' thought - and in the qualitative lexicon in general - than it does in the quantitative scientific tradition. Theory-making in qualitative social research is seen less as an effort to identify and nail down the immutable laws of the universe, as if the world were simply a vast unchanging clockwork to be discovered piecemeal, than as an effort to construct alternate perspectives that throw new light on inherently rich and complex situations and help us to better understand their dynamics. In the qualitative tradition, theory is essentially what the Greek philosophers called a "heuristic" - that is, an aid to discovery, a learning tool.

The issue of "objectivity"

Qualitative perspectives are not always accepted in the bureaucracies and organizations that determine educational policy. One of the first and most persistent problems that qualitative evaluators run into is the criticism of subjectivity. Decision-makers want "objective" data, not fuzzy impressions. (George Stigler, the Nobel prize laureate in economics, is reported to have quipped, "Anthropologists? Aren't they the people who think the plural of 'anecdote' is 'data?'") But what is meant by "objectivity"?

Educational programs naturally develop their own ideology. Political necessity may surround them with a thick coat of rhetoric and dictate that those dearest to politicians' hearts be reported and even lionized as successes whatever their actual results have been. Penetrating these appearances and establishing a new and healthier discourse about social programs may be one of the most critical functions of evaluation. At one level, it is true that evaluation is precisely about objectivity, about getting what might be called the unvarnished facts and discovering what is "really" going on as opposed to what is popularly presented or thought, about gauging the actual effects of programs as opposed to their reputed or advertised ones.

But does this objectivity necessarily or exclusively mean numbers? Etymologically speaking, the word "data" is the plural of "datum," Latin for a single "given" or fact; and the word "fact" refers to "a thing done" (*factum* in Latin, past participle of the verb *facere*, "to do"). Most of what are popularly called facts, however, are really aggregates, construc-

tions, or compounds of such single elements or observations. When we say there are thirty people on the football field, as illustrated earlier, we include in that observation a whole set of definitions and concepts. The "fact" that a certain company went bankrupt is actually a characterization of a series of events and a particular categorization of financial conditions. In addition, from a certain point of view, one could even claim that qualitative methods are more "factual" than quantitative ones, because they typically involve observing and describing events and settings in much greater detail and with richer nuance than do quantitative methods. Yet whoever perceives them, these observations of experience are not simply neutral transcriptions of the reality of the universe. Qualitative researchers consequently speak of the "theory ladenness" of data - i.e., the degree to which facts are in fact a product of conceptual presuppositions and so are replete with theory from the outset.

Individual "facts" are much too numerous and various to hold in consciousness, rather like grains of sand or snowflakes. In truth, it is not even clear what "facts" are, since identifying discrete elements in the flow of experience requires subdivision and segmentation that is always influenced by our prior beliefs outlooks. We *do* need patterns to help us organize and filter the flow of experience, and to distinguish, store and later use parts of it. But these patterns never function perfectly: sometimes it becomes apparent that things have been left out, at other times we find our frameworks are adapted to one set of circumstances but not to another. We need continually to modify and reformulate our ways of viewing the world, both as individuals and as collectivities of people.

Qualitative evaluation helps us to reformulate our ways of seeing programs and social institutions. As an exercise, it helps us to become more conscious of when and how our frameworks and the categories of our understanding are screening out critical events and phenomena with which we need to deal. This sort of evaluation provides an opportunity to try out new frameworks, gather new facts and patterns, and modify the mindset and model of a social program in ways that will make it more effective.

The basic impulse of qualitative evaluation is therefore to go back to the well of experience and social meaning from which programs are built, back to the flow of life from which selected facts have been drawn and to the sources of patterns for constructing and interpreting them, in order to help build up a new vision of what the program is and should be doing, or to help modify the existing one sufficiently to make it more

serviceable. It is moreover precisely because one goes back to the constituent elements, the "facts," that (alternate) frameworks are so important. Indeed, it seems fair to say that this sort of exercise in renewed objectivity *requires* qualitative tools that help us perceive dimensions and elements of our environment that were not apprehended before or that are passed over in current popular images of the situation. But we must first break our present blinders - our current meaning patterns - and develop or entertain enough new ones to allow us to see what we missed before.

To summarize, objectivity means returning to the facts, or as many of them as we can capture with our present technology and perceptive competence. But these facts are only accessible via some method of framing experience, through some prior perspective. To recognize and capture *new* facts, we must frame our inquiry in *new* ways, then see how to build our perceptions up into new patterns of understanding, new insight. So objectivity by definition requires *both* qualitative and quantitative procedure. Qualitative gives us exercise in developing, discovering, recognizing patterns, building them up - and exercise in clearing ideas of ideology: i.e., trying to separate insights from class-bound, gender-bound and status-bound bias. Quantitative provides a basis for verifying patterns over the large population groups that typically are involved in social and educational programs and for specifying inputs, processes and results in ways that are easily comparable and accountable. Evaluators need by definition to walk on two legs.

We turn now to some of the different kinds of qualitative investigation that may be used in the course of an NFE program evaluation. The techniques most frequently used in actually carrying them out will be considered in Chapter Nine.

Types of qualitative inquiry

Qualitative elements within an evaluation study may take many forms. Some of the most frequently encountered are the following, and each has a number of sub-species:

1. *Case studies* of individual programs, sites, groups or persons. In a case study, one focuses on careful description and understanding of the particular entity under consideration - its history, its context, the

nature of its activities, its relations with other things in the environment, and particularly how it operates in everyday life. The notion is that by carefully and empathetically examining this particular instance, one may become sensitive to heretofore hidden or poorly understood aspects of many other cases as well. (cf. Hamel 1993; Yin 1988)

2. *In-depth interviewing* of stakeholders, whether individually or in groups. The length and nature of these interviews may vary greatly - from open-ended questions in otherwise quantitative surveys up to extended biographical interviewing where one is helping the interviewee(s) to reconstruct and reflect on significant portions of personal life history or of the development of the program being evaluated. Group interviews, it should be noted, create quite different dynamics from individual ones. Cultural mores regarding who talks to whom, under what circumstances and in what ways need to be carefully taken into account in constructing interviewing situations. (cf. Holstein 1995; Kumar 1989)
3. *Participant observation* in its many forms is one of the most important categories of qualitative research and evaluation (cf. Bogdan & Biklen 1992; Patton 1990). It involves the researchers or evaluators entering directly and personally into the situations under study, witnessing what transpires and taking part in these events on a basis at least approximating that of local participants in order to "record" mentally and/or materially (e.g., via notebooks, tape recorders, video or still cameras) as much of the scenario as they can and as they find relevant. The degree to which the evaluator or researcher actually becomes a member of the group in question and undergoes the same experiences that others do is variable. It might be very worthwhile, for example, that an evaluator experience just what it means to work as a newly literate paramedical aid in a village infirmary, but it is doubtful that he or she could duplicate by mere presence the actual experience of *being* newly literate and learning certain medical notions for the first time as an adult.

Two criteria generally applied to decisions about the role of participant observers are, first, how to gain maximum feasible empathy for and understanding of what program participants or stakeholders experience and, second, how to minimize the "obtrusiveness" of the researcher's presence. One is generally interested in experiencing the events or activities in question just as they

would "normally" take place in the absence of outside observation. In some circumstances, outsiders will be least obtrusive by mimicking the behavior of group members and, when in Rome, "doing as the Romans do." According to a West African proverb, if you discover that the village chief has grown a tail, it is high time to go to the field, chop down a corn stalk and tack it to your behind! In other circumstances, the difference in culture or competence is so great that an attempt to behave as if one were an insider - like an urban researcher trying to participate in agricultural life by using potentially dangerous farm equipment - would be more disruptive than discreet observation.

4. *Documentary or "artifact" content analysis* entails careful examination of the many different types of materials produced by participant or stakeholder groups in direct or indirect relation to the program (cf. Krippendorff 1980). These include notebooks, administrative manuals and memoranda, other publications, drawings, oral traditions, tools, furnishings, buildings and virtually all the trappings of the "made environment." Any and all of these products of local culture may be very expressive of how people view and value the program or phenomenon in question. Just the way in which a training center, school or literacy headquarters is located, constructed and furnished may "speak volumes" about its place in local society. Note that these artifacts may be described both qualitatively and quantitatively.
5. *Mixtures of these categories*: Case studies obviously include large segments of in-depth interviewing and artifact analysis. Participant observation will be richer if it is undertaken on the basis of the kind of understanding that is provided by good case studies. Most actual qualitative evaluations combine several variants of the methodologies suggested above.

In addition, the strategy used in any study will vary according to the sampling methodology adopted to identify the people, places or events to be studied. One may, for example, be looking for exceptional cases that seem to demonstrate what a program could become, or for patently average ones, or for notable failures. It all depends on the questions one wishes to address.

Research design considerations

The purpose of research or evaluation design is to organize inquiry in such a manner that we ask the right questions, obtain the type of information that will help us to answer those questions and analyze or process it in a way that will provide enlightening answers. Quantitative designs, as discussed in earlier chapters, entail specification of variables and indicators, their operationalization through instruments, and careful planning for the kind of analysis to which the data will be subjected. The weakness of such an approach is that theory - with all its built-in blind spots - is put on up front. Enumeration is deeply imbued with prior assumptions, so if our objective is to renew and broaden our perspectives, we are already "behind the eight ball."

In qualitative designs, by contrast, the principal "instrument" is the set of persons involved in the research or evaluation project: the researchers or evaluators, the interviewees, and the authors or artisans of the documents and artifacts used in the inquiry. Qualitative approaches capitalize on the human facility for perceiving pattern and discerning meaning. Rather than being defined rigidly in advance, variables, indicators and theories *emerge* from the inquiry as the researchers gain an appreciation of what they and stakeholders see as key factors. Many more choices are made in the field.⁷ This means that a higher level of competence is typically required in qualitative inquiry than in the execution of fully pre-established designs.

Similarly, as we saw in Chapter Seven, sampling takes on a new meaning: one is not looking strictly for demographically representative individuals or cases, but rather for those that are particularly *rich in information or potential insight into what is really happening with the program*. (See in this regard the variants outlined in the previous chapter and the example presented in Box 8.3 hereafter.) Key informants, particularly graphic cases, and "limit examples" - i.e., exaggerated situations that illustrate particular dimensions of a problem or an important latent potential - become critical, as well as all those especially relevant to the central questions of the inquiry.

There is an additional danger of bias here, of course: the most articulate stakeholders and the most remarkable cases usually constitute a peculiar subgroup. As in all research designs, whether quantitative or qualitative, one is faced with a tradeoff. The point is to identify particularly informative cases and to "plumb" them in a way that gets the full

*Box 8.3. Elite sampling in Morocco:
Picking atypical situations*

Qualitative researchers know full well that "extreme" situations often reveal the dynamics of everyday life better than does observation of more routine events. The roles that people play, the values that they hold and the most fundamental relationships that link them may be more transparent under circumstances of stress or change than they are ordinarily. In a similar vein, Marxist social scientists have often maintained that the best way to understand a social situation is to try to change it - precisely because the stimulus or threat of change typically brings "out of the woodwork," so to speak, stakeholders and interests that otherwise remain hidden.

Evaluators frequently need to pay heed to the same principle, which is one reason why they are generally admonished to attend to the "glorious failures" of a program as well as to its success stories. During an evaluation of rural literacy and vocational training programs conducted in Morocco in 1994, it was in those centers where nothing resembling the official program had transpired that evaluators found the clearest expression of the resistance that the methodology had provoked in certain quarters and of the needs it had failed to meet. Careful selection of places where the shortcomings of a program (and every program has them) are most clearly displayed is itself an important form of "elite" sampling.

richness of insight that they embody or can articulate, while simultaneously controlling for overt bias. The best such control generally comes from triangulation of different points of view.

Finally, since much of qualitative design must take place in the field, it necessarily has an artistic and intuitive side, if not the appearance of a haphazard one. The evaluator never knows exactly what he or she will find and must be ready to put whatever the context reveals to good use. One good metaphor might be the art of Japanese painting, in which both the intentional figures drawn by the artist and the ink spots accidentally spilled by her child on the paper are taken to be a significant part of the work and blended in an inspiration that expresses the beauty and meaning of that moment.

The chance occurrence, the unforeseen obstacle, the fortuitous encounter all are thus full of meaning for qualitative investigators; these events are often what help them get outside of their own unconscious

frames of reference and expectations - or at least to become more fully aware of them!

Conclusion: The strength of soft designs

Qualitative designs may appear more fluid, not as "hard" as quantitative ones, given their lesser concern with statistical enumeration and with standardization; but they are arguably at least as strong.⁸ One of the underlying themes of any research or evaluation endeavor is to capture and analyze the *variability* in program processes and outcomes, for understanding variability and explaining differences are key means of analyzing causality and the true nature of relationships - in short, the underlying dynamics of an activity. Qualitative researchers actually take variability a good deal more seriously and systematically than do quantitative ones. They amplify and describe it in greater detail. Changes in data that result from examining different aspects of a situation and from viewing them via different approaches, at different times, and from different perspectives are carefully cultivated and captured. "Multiple realities" are investigated - that is, the varying meanings that different people assign to the same events or circumstances. Triangulation, or the comparison of varying views of data, is nearly the *leitmotif* of qualitative methodology.

In the next chapter, we will take a closer look at the *how* of qualitative evaluation, particularly the techniques and artistry involved in collecting data in the field, progressively analyzing it, and constructing its meaning. But the reader should have by now a better grasp of the claims of qualitative evaluation to validity and of the important role that it plays in an overall evaluation effort or "macro design." This chapter and the following one span two sections of the book, because in passing from design considerations in qualitative evaluation to actual field work, we also move from the predominantly conceptual portion of our topic to a certain number of practical methodologies that are important in evaluation work and that receive less attention than they are due.

Notes

¹ In fact, policy analysts in Congress often referred to this situation as the problem of "horse and rabbit stew" - it doesn't take much horse meat in the mixture to dominate the taste and overpower the other element. So as soon as quantitative data are introduced into a policy report, they felt, all attention were focused on the numbers and other arguments or evidence were forgotten.

² Aaron Wildavsky humorously and trenchantly subtitled his well-known book on the implementation of public programs during the federally-mandated War on Poverty in the United States: "How great expectations in Washington are dashed in Oakland; or, Why it's amazing that Federal programs work at all." (Wildavsky 1973). Oakland, California, it should be noted, is 3000 miles from Washington, DC., the United States capital city... in more ways than one.

³ For guidelines on historical inquiry as a method of qualitative research, see Berg 1995 and Richardson 1986.

⁴ One commonplace application of this latter principle is well known: People generally have better luck remembering lists of names or random numbers if they can connect them with what psychologists called a "mnemonic device" - that is, some familiar association that is easy to recall. For example, if a basketball fan notes that the last four digits of a friend's telephone number are the same as the score by which the local team recently beat an arch rival (e.g. 89-75), the task of remembering what to dial is likely to be much less difficult.

⁵ I am indebted to Dr. George Papagiannis for the following illustration.

⁶ This example makes it evident that the notion of "meaning pattern" presented here is quite close to that of a theoretical "paradigm" first popularized by Kuhn (1963) as a way of describing the structure of assumptions on which scientific knowledge is built. The term "paradigm" has been used much more widely if not abusively in recent years to characterize everything from different styles of business organization to different evaluation styles.

⁷ Many but, of course, not all: qualitative researchers are scarcely exempt from bias and preconceptions themselves, but they do make a concerted effort to recognize and analyze these before and during their fieldwork.

⁸ This is not to say that numbers are never used in qualitative evaluation and research. Quite the contrary: numerical data may be very useful in describing the environment and probing certain characteristics of the cases being studied. In a qualitative evaluation of learning patterns among new small

business owners, one will not want to omit data on the size and financial health of those organizations or on some of the attributes of their customers. Descriptive statistics are therefore often prominent parts of qualitative studies. The difference is that the main thrust of the research is not to analyze statistical relations among quantitative indicators of a situation, but rather to understand its richness and variability and the meanings that interested parties attribute to it. Both qualitative and quantitative evaluations use numbers and descriptive statistics, but quantitative ones specialize in inferential statistical analysis.

References

- Berg, B.L. 1995. *Qualitative Research Methods for the Social Sciences*. 2nd ed. Boston; London : Allyn and Bacon.
- Bogdan, R. and S. Biklen. 1992. *Qualitative Research for Education: An Introduction to Theory and Methods*. 2nd ed. Boston: Allyn and Bacon.
- Glaser, B. and A. Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine Press.
- Hamel, J. 1993. *Case Study Methods*. Newbury Park, CA: Sage Publications.
- Holstein, J.A. 1995. *The Active Interview*. Thousand Oaks: Sage Publications.
- Krippendorff, K. 1980. *Content Analysis: An Introduction to its Methodology*. Beverly Hills: Sage Publications.
- Kumar, K. 1989. *Conducting Key Informant Interviews in Developing Countries*. Washington D.C.: Agency for International Development.
- Patton, M.Q. 1990. *Qualitative Evaluation and Research Methods*. Newbury Park, CA: Sage Publications.
- Richardson, J. 1986. *The Local Historian's Encyclopedia*. 2nd ed. New Barnet: Historical Publications Ltd.
- Yin, R. 1994. *Case Study Research: Design and Methods*. Thousand Oaks, CA: Sage Publications. 2nd edition.

Part IV
**Particular Methodologies
of Evaluation**

CHAPTER 9

THE PRACTICE OF QUALITATIVE EVALUATION: ARTISTRY AND TECHNIQUE

This chapter is devoted to a brief examination of techniques for performing qualitative evaluation in the field. The word "technique" is used both advisedly and designedly: advisedly, because qualitative evaluation is as much an art as a technical procedure; designedly, because the craft clearly does have its techniques, and policymakers (or researchers) of a quantitative bent need to be reminded that good qualitative studies are at least as exacting and systematic, in their own way, as are quantitative ones. In the pages to follow, we shall look successively at approaches to collecting and managing qualitative data, to analyzing the material thus assembled, and to combining it with the quantitative results of an evaluation.¹

Collecting and managing data in qualitative designs

As instruments for gathering data, analyzing that information, and using the results to orient further inquiry, human participants obviously have much more potential than do preconceived tools like questionnaires. They also pose, however, heightened problems of "reactivity". Biases and personal interactions may come into play much more prominently, at least in the field portion of research, when evaluators are gathering data from their direct experience rather than collecting existing numbers or soliciting responses to standardized instruments. (As noted in the last chapter, biases in those quantitative procedures tend to be put on up front in the definition of data and the conception of instruments.) Intrusiveness likewise becomes an issue: the presence of the researcher or evaluator may influence results. The art of fieldwork lies in maximizing the richness and minimizing the dangers of this method of data collection, in order to get the most from the "human instrument". We will discuss how this is done under five rubrics: controlling bias, obtaining entry, transcribing data, managing

the divergent and convergent phases of fieldwork, and facing ethical dilemmas.

Controlling bias

People are the prime instruments of qualitative inquiry, but they are a highly complex kind of instrument, and their own viewpoints inevitably interact with what they see. That is not necessarily bad. In fact, it may be quite good, *if one knows how to take account of it*. Taking account of it usually means making a systematic effort to explore and articulate the biases, or the particular lived viewpoints, of those involved. Major qualitative studies often include a special opening section or an appendix where the researcher tries to examine his or her own previous experience with the situation under investigation, his or her feelings about it, and the aspects of the researcher's own background and culture (social or class position, ethnicity, beliefs, etc.) that may influence how he or she perceives and is perceived in the situation.

Such a full-blown exploration of potential bias or personal viewpoint may not be appropriate in every qualitative dimension of an evaluation. Still the *principle* that evaluators should attempt to take systematic account of the aspects of their own "persona" and belief systems that may interact with the subject of their study remains valid and is worth noting.

Obtaining "entry"

How does one manage to see the right people, to observe the key events and situations, to identify and examine the most telling documents or "artifacts" of a program? Qualitative evaluation focuses on "in-depth" information of a kind not usually available at first glance to the quantitative surveyor or to the casual passer-by. But this presupposes *access* to such "privileged" information, to reactions and insights not always verbalized, and to settings that may be removed from the public eye. The banal is always easier to locate than the profound. And of course, most subcultures and organizations erect invisible barriers around themselves - be they high or low, relatively porous or more impenetrable - in order to provide protection against intrusions and

information leaks that members see as potentially threatening to their security or social cohesion.² For these reasons, qualitative researchers typically devote a good deal of attention to the issue of "entry" - that is, the means by which they will secure admittance to the settings they wish to study or evaluate, gain the confidence of their interlocutors, and succeed in locating and learning from information-rich cases. One method is obviously to spend an extended period living in the environments under study. Ethnographers often try to become intimately familiar with the "cultures" they are studying through prolonged exposure and close contact.

Not all researchers are at liberty, however, to spend months or years living in the field; evaluators very seldom are. To the degree that direct exposure to the study environment is limited, a second method of "entry" becomes important: use of "key informants" or guides. These are people "on the inside," native to the environment in question or highly familiar with it, who can help researchers or evaluators to identify and gain access to information-rich situations and "elite interviewees" and help them to interpret what they see and hear. The choice of such people and the cultivation of appropriate and honest relationships with them is obviously fraught with some perils, both because, as a stranger to the milieu, the evaluator may choose the *wrong* person as confidant, and because this sort of rapport with outside "investigators" may cast the local counterpart in a poor light within the community. There is in fact a whole literature in anthropology on the relationship between researchers and key informants.

Evaluators face special barriers in this area if the evaluation itself is perceived as potentially threatening to certain stakeholder groups or entrenched interests. Yet this problem may be considerably lessened in a genuinely participatory evaluation, since in such cases local stakeholders themselves become guides and interpreters by virtue of the important role they play in designing the evaluation, collecting the data, and analyzing results. It is moreover largely *their* responsibility under these circumstances to interpret the results of their own local assessment in such a way that this information will be understandable to other communities, as well as to outside evaluators responsible for facilitating and coordinating the process.

Whatever the degree of stakeholder participation, however, it is evident that in qualitative evaluation gaining entry to relevant contexts

and people is very important and that considerable thought must go into identifying the program elements to observe, the people to interview in open-ended ways, and the questions to ask them (for good questions are like key entry points to the mind). Since one can never sample as broadly in a qualitative inquiry as one can in a quantitative study, it becomes proportionately more critical to pick those sources of information that will provide the richest insight.

Transcribing data

One of the prime technical problems in qualitative evaluation concerns how to note down or transcribe observations in a way that ensures the "thickness" of thick description. When one is filling out questionnaires, there are boxes to check off; but when interviewing is open-ended, it is less clear what to write down. One cannot note everything the way a court stenographer does: few people have such talents, and the very dexterity involved would probably be distracting for the interviewees. More importantly, word-for-word transcription may yield uninterpretable mountains of data. Similar problems plague the tape recorder solution. Taped sessions are nightmarish to transcribe, since fifteen minutes of speaking can easily require several hours of transcription work, and the volume of output is very difficult to deal with thereafter.³ On the other hand, highly summary or schematic procedures, like making brief notes on the interview the same evening, forfeit most of the sap and substance that are the *raison d'être* of qualitative inquiry. Why go to the trouble of identifying and gaining access to information-rich cases if one derives from them little more than one would from a standard questionnaire?

Some compromise between the two ends of the spectrum must be sought, particularly in the case of qualitative *evaluation* where the resources available for full-blown phenomenological or ethnographic inquiry are rarely the same as they are in research. That compromise generally involves moderating or modifying transcription methods in order to make them more *selective*. Selecting the "most important things" is of course a judgment call, but the admission simply means that data analysis has already begun. Ample enough descriptive detail must also be included to place the data firmly in context without drowning perception in detail. It is easier to edit intake than to edit

output, so the qualitative evaluator has an interest in refining as much as possible his or her internal sampling of events, people and topics in such a manner as to avoid redundancy and hit the high points. Remember that some of the most fundamental issues in an evaluation do not require massive samples, but rather good discernment.

Here are some illustrative techniques and tricks of the trade that may be useful in focusing transcription tasks:

- *The use of teams* of two or more qualitative evaluators often allows one team member to conduct interviews or group sessions while the other(s) take(s) notes. Good note-taking is an art that is always worth practicing and improving, because so much of data transcription in the qualitative mode comes down to this.
- It is advisable to *plan for time out* after each observation or interview session in order to allow the evaluator to note down immediately, while impressions are still fresh, a synopsis of what was observed or heard, with as many relevant direct citations or descriptions as possible - or, if such primary data transcription has already been handled by other means, to add some notes on personal perspective or interpretation. Practically speaking, this generally means (gracefully) ending the interview or exiting the event before one's time and energy are both spent - a habit that requires some self-confidence, some self-discipline and a good sense of timing.
- *Complete tape recording* of events and interviews may be useful even if one does not intend later to transcribe the entire proceedings. The tape can be considered a reference and backup that gives evaluators a way to check passages that were unclear or difficult to remember. Evaluators may also choose to listen to an entire sequence a second time through, taking notes as if they were once again in the situation. (Moreover, to the extent that this method is practiced, evaluators may choose during the initial session to take notes particularly on the "inaudible" dimensions of the situation and setting that won't later come across on the tape: gestures, expressions, staging, furnishings.) Comparison of the first and second sets of notes will considerably enrich the data. One may also "sample" sections from the tape for fuller (content, linguistic) analysis and/or catalogue contents with a counter.

- "*Member checks*" is a term used by ethnographers and qualitative researchers to refer to the technique of verifying the contents of interviews with the interviewees once these encounters have been transcribed and analyzed. There is both an ethical and a technical reason for doing this. The ethical one concerns not using someone's words - particularly in regard to the sometimes delicate issues raised in program evaluations - without his or her consent. Interviewees should have an opportunity to clarify and even edit what they have said, though of course the evaluator will at one level remember the differences between the first and second versions of their comments. The technical reason for member checks is simply to correct errors of transcription or of interpretation that may have crept in. Such second-round contact may provide an opportunity for going even deeper into the topics covered and may therefore lead to new and more complete data. In participatory evaluation, the notion of "member checks" naturally expands to include much more highly developed forms of consultation and joint analysis of data with stakeholder representatives. In *any* qualitative evaluation, however, some formative feedback to key informants or interviewees and some solicitation of their reactions is highly appropriate. Qualitative evaluators are more like co-researchers with their local counterparts than they are like laboratory scientists.⁴
- *Notation of feelings, reactions and subjective judgments*: Qualitative researchers take care to note along with their data, and often directly in the margins of their notebook or other data transcription device, their own personal reactions to what they have observed and their interim judgments concerning it.⁵ Such subjective material is considered to be part and parcel of the data. This technique can be very helpful for two apparently contrary reasons: first, because potential bias should always be taken into account and flushed into the open, and this technique helps to separate it from the transcription of observations; and, second, because these feelings may nonetheless be highly accurate and provide some real insight into what is going on beneath the surface of the social events observed. Noting reactions and feelings is another means of focusing the inquiry. It helps the evaluator to draw interim conclusions about the relevance of different types of data collected

and therefore to make better choices regarding those still to be sought.

- *Writing, writing, writing:* During qualitative data collection, several "levels" of notation are going on at once: primary transcription or description of phenomena, notation of feelings about them or reactions to them, and analysis of relations among them. Qualitative researchers make a habit of writing memos to themselves and treating these materials as data as well. In fact, good descriptive writing is arguably the fundamental "technology" of qualitative research, since notes and written observations and reflections are such a major part of the data collected.

Managing divergent and convergent phases

These tasks of focusing and editing the transcription of events simply reveal that data analysis is beginning already. As we shall see in a moment, analysis in qualitative methodology is not something left to the end of inquiry, but occurs throughout the experience. One is constantly sorting and defining data and making decisions about what kinds to collect next - an additional reason why qualitative evaluators must be trained well and then left to their own devices to a greater extent than is typically the case in quantitative research.

Managing the divergent and convergent phases of data collection so that one leaves the field (whether that "field" is a group of villages, an urban neighborhood, a factory or an administrative office) with a reasonable quantity but not a crushing amount, of well selected, richly described and partly processed qualitative information is therefore the challenge. Meeting it requires an aptitude both for ferreting out and for capturing significant experiences and for focusing and selecting data input from them. There are in fact rhythms in any participant observation or qualitative research endeavor, from entry through intensive note-taking to preliminary synthesis and more directed inquiry, to late-stage organization and systematization. The early stages are by nature more *divergent*, the later more *convergent*; but, ironically enough, qualitative evaluators are impelled throughout to do at least as much regular and systematic (if non-probabilistic) sampling as do most quantitative evaluators, because they can so easily be overwhelmed by the amount of potentially important data and are typically more sensi-

tive to the dangers of excluding relevant material. Qualitative evaluators will not and should not stick with just those potential sources of information that they have identified at the outset and must remain open to multiple types of new data, including notation of and reflection on some of the apparently unconnected or "extra-curricular" activities in which they become involved while in the field. But by this same token, great care must be exercised to avoid collecting a glut of unusable notes and information.

Finally, we should note that data transcription does not mean simply producing written summaries of what has been observed or heard. To begin with, it is important to capture as much of the story as possible in the participants' own words - which, in the majority of cases where complete transcription is not the method used, means noting down selected excerpts of interviewees' discourse, particularly pungent phrases or eloquent characterizations that they may have used. In addition, there are many other kinds of data that are worth "transcribing." Artifacts produced by program beneficiaries themselves (their notebooks, letters, newspapers, crafts, etc.), as well as drawings, photographs and videotapes of the setting and the events will greatly enrich the data base of the evaluation and the possibilities of analysis and interpretation.

Ethical dilemmas

When one seeks to penetrate this deeply into what is going on and why, questions of personal feeling and political vulnerability are soon involved. We must therefore always think through what we are asking people to do or say and to what distress or danger that may expose them. Issues of privacy, anonymity, pseudonyms and attribution inevitably arise and must be resolved by evaluators in a way that is consistent with a deep respect for human rights. (Box 9.1 presents a pertinent example.)

Here again, participatory approaches to evaluation, while they have may uncover controversial material at times, can also serve to attenuate the problems posed by the potential vulnerability of respondents. Since data production and debate are out in the open and there is a collective commitment to conclusions, individuals may be less exposed to the dangers of too much candor than in an individual-

interview or outside evaluator-driven procedure. When everyone is saying something, it is less easy to pick scapegoats.

But in any and all cases, qualitative evaluators must remember that the prime purpose of their work is to benefit program participants and that the rights and needs of those people should therefore be the highest priority criterion in evaluation design and implementation decisions. (For a good analysis of ethical issues in qualitative research, see Punch 1986.)

Data analysis

In qualitative evaluation, we are essentially attempting to discover the "meaning patterns" that explain why people do what they do, how they interpret events, and what roles they play in an educational program.

*Box 9.1. Ethical dilemmas in evaluation:
The "bèe be min don" effect in Mali*

Though evaluators in theory want "the truth and nothing but the truth" about program operations, there is a saying in French that is equally applicable to their work: "Not all truths are good to speak." Some topics - and often those concerning use of resources or the performance of politically significant people - are so controversial or potentially explosive that evaluators would be risking jeopardy to their respondents and counterparts, if not to themselves, in broaching them too openly or attributing comments about them to other parties. The unvarnished statistical results of a program may have this sort of explosive potential, particularly if they are well below what was previously expected or claimed; but the problem most frequently arises in qualitative evaluation, when one is probing in greater depth just what did indeed happen and why.

This was the case during a literacy evaluation conducted in the Republic of Mali under the former military government, and such topics were frequently referred to by the evaluation team as the "bèe be min don" factor: a Bambara-language expression for "the thing that everyone knows" (but that no one dares say). Evaluators must demonstrate a great deal of perspicacity and tact in deciding how far to venture forth on such terrain, how veiled to make any references to related subjects, and what precautions must be taken to protect their key informants.

How does the evaluator derive this kind of insight out of the mounds of data that may be collected during a qualitative study? How do patterns emerge from what may be great quantities of verbiage and observational notes.

Qualitative *evaluation* is generally less exhaustive in data collection than qualitative *research*, and analysis problems may be correspondingly less daunting; but a cautionary word from two practitioners is quite pertinent and will serve to introduce the technical remarks to follow:

The data generated by qualitative methods are voluminous. I have found no way of preparing students for the sheer massive volumes of information with which they will find themselves confronted when data collection has ended. Sitting down to make sense out of pages of interviews and whole files of field notes can be overwhelming (Patton 1980: 297).

Data analysis is the process of bringing order, structure and meaning to the mass of collected data. It is a messy, ambiguous, time-consuming, creative and fascinating process. It does not proceed in a linear fashion; it is not neat. Qualitative data analysis is a search for general statements about relationships among categories of data; it builds grounded theory (Erlandson et al 1994: 111).

Two approaches

Qualitative evaluators and researchers take two different though fundamentally complementary approaches to the task of data "reduction" and analysis. Eisner (1990) has dubbed these the , "scientific" and the "artistic" orientations. The former is characterized by an effort to disaggregate qualitative data into its component bits of text or observation and then to code these in order to regroup them and build them back up into patterns that explain events and behaviors. The "artistic" approach, on the other hand, is more holistic: it consists essentially in an effort to grasp the "flow" of events or the internal coherence of people's words and activities and to express these. Eisner also calls this latter method "educational connoisseurship" because it puts a premium on the insight of the evaluator and on his or her ability to discern the meaning in events and the underlying qualities of a social

setting from sense observations, just as a "connoisseur" of, say, fine wine can tell from a taste where the grapes were grown, how they were processed and how well they have been aged.

I will treat these two varieties of qualitative analysis, however, as two steps in a common process. The "scientific" evaluator places more emphasis on the divergent coding and categorizing task, whereas the "artistic" evaluator stresses the convergent or synthetic challenge of making meaning from sense observation; but both necessarily accomplish the two tasks in some form. Arguably, the common denominator method used by both is what Glaser and Strauss (1966) first called the technique of "constant comparison": that is, repeatedly comparing words and actions, events and the physical characteristics of settings in order to discover the ways in which they are related and the meanings that people assign to different aspects of a program; then comparing new data to these derived patterns in order to confirm, modify or disconfirm them.

In both styles of analysis, moreover, this effort is not confined to the moment when the evaluator leaves the field to process the data collected. If there is one characteristic of qualitative evaluation and research that all practitioners stress, it is that *data analysis necessarily begins in the field*. The qualitative evaluator must sift, organize and interpret what he or she observes in the very process of collecting data, refocusing the study and better defining its themes in response to what is discovered. Early results of this nature guide later decisions about what further data to collect - i.e., what other people to interview, what additional events to observe, etc. In this section, we will consider how the task of ongoing discrimination and synthesis is accomplished.

Steps in deciphering meaning

In qualitative inquiry, we attempt to reveal and express those patterns that are implicit in local culture (whether it is the culture of an office, a school or a village), in people's actions and in their accounts of what is happening and why. Sometimes this meaning can be relatively easily apprehended. At other times, very close attention is required to "hear" what is really being said, to perceive implicit rules about who can do what and how, and to understand what local counterparts see as the

likely causes and effects of social action. The central data analysis technique in qualitative evaluation is simply a particularly sensitive and systematic way of "listening" or "attending" to what is happening, of seeking the inherent "codes" in communication and observation - that is, the themes that come back or the events that reoccur, the connections in which they do so, and the things left unsaid because they are so obvious to insiders - in order to begin discovering a whole underlying pattern of relationship in what people do and in how they understand what they do.

How do we identify and describe such patterns? From notes on what is seen, what is done and what is said: what people express in words, and what they express in actions, configurations, postures, and all the non-verbal manifestations of attitude and meaning. Qualitative data are like the kind of multifaceted evidence that might be presented in a courtroom or consulted by an investigative journalist. We are in effect listening to what people say in their silences and their daily itineraries as well as in their words. One may need to know several languages to capture what is really happening in any given setting: formal languages (e.g., French, Cantonese, Yoruba), informal linguistic codes (like the differences between men's and women's speech), other subcultural variants in expression, perhaps technical languages like those of agriculture or auto mechanics, plus the artistic or representational ones used in popular music, painting and architecture.

Qualitative data analysis is an exercise in *deciphering* these various codes in order to discover "what is really being said" - that is, the underlying dynamics of social situations and the meanings that people attribute to them. (Box 9.2 provides a pertinent example.) Techniques for deciphering are many, but the most frequently-used approach consists of regrouping these data according to a set of categories or themes that emerge from an attentive examination of them, next trying to summarize the evidence with regard to each topic or theme, and finally looking for patterns and relationships that appear between the different topics. The steps in the process might be summarized as a series of five phases: discovering themes and categories; coding and grouping the data; analyzing and understanding the data; synthesizing these insights; and validating the results. They are in fact more interwoven and overlapping than such a sequence would imply, but those five rubrics provide a framework for discussing the topic.

*Box 9.2 The payoff to good listening:
Local pronunciations of the word "coopérative" in Mali*

Qualitative evaluators investigating the uses of literacy in western Mali in the 1970s spent hours discussing the history of the farm cooperative movement with groups of villagers. Though local experience with government-run cooperatives had reportedly been turbulent in many cases, farmers were at first quite reserved and circumspect in their remarks. Listening carefully though, one evaluator noted that the Bambara-language expression used for the subject in their remarks was not simply a fractured rendering of the French "coopérative", but a poignant pun. They were in fact saying "ko-fara-tinti" which in Bambara signifies "skin" (*fara*) the "back" (*ko*) and "skewer" or "stab" (*tinti*): in short, an exercise in getting fleeced and backstabbed. The evaluator was then able gradually to guide interviews toward the question of how local people perceived that they had been "skinned" and "skewered" by the former cooperative movement - and what this meant for future attempts to organize locally-managed cooperative markets as an application of new literate competence.

Discovering the themes and categories

The work of "listening" and understanding starts during field research, which is why we remarked earlier that data collection and analysis are usually interwoven in qualitative inquiry. The concern is to discern the primary themes to which people's actions and words relate, the principal categories in *their* perception of the program and of the areas of social life that it affects. (Box 9.3 illustrates the discovery of one of these implicit themes.)

Later on in the analysis, what is *not* said, categories that do *not* appear, may be as important as those that do. But for the time being we are interested in grouping the data brought back from the field around the recurring themes and daily topics that underlie this domain of social life. Once again, these derive as much from observation as from the spoken or written word. In many West African countries, for example, the location chosen for community-built training centers, the groups responsible for their construction and maintenance, and the general state of repair of these facilities may say volumes about the degree to which the training activity has been truly adopted and sponsored by local leaders before a word is spoken or the first inter-

BEST COPY AVAILABLE

view conducted. Some of these categories will flow quite naturally from the different domains and aspects of the program itself enumerated in program theory and implemented in the field; others will concern the local institutions, habits, events and actors involved.

There is reason to make the list *relatively* parsimonious: fewer than one hundred categories, perhaps, which may then be regrouped under a smaller number of overarching codes. The idea in regrouping things is to associate like elements that may throw light on each other by creating distinctly fewer "slots" than there are items and trying to divine "common denominator" issues. Nonetheless, the list should certainly remain open, because rereading of the data or inclusion of new information may suggest additional categories or reformulation of

Box 9.3. What's in a school?

Definitions of educational investment in Haiti

Qualitative economic researchers (yes, such creatures exist!), who were investigating how families in the low income neighborhoods of Haiti's urban centers make budget decisions about education in a context where most schooling is private, discovered that people implicitly categorized the costs of schooling as critical *consumption* expenditures more than investment ones, even though they were quite accustomed to making small investment decisions on a regular basis (e.g., purchase of buckets to carry and sell water). The reasons were simple, though not always evident to national planners who justified education projects in the country on a "human capital" rationale.

First, returns to education, while real, were far below those that could be made in small scale commerce or loan operations with even very minute amounts of discretionary capital. Second, most local schooling at that time served simultaneously as a conduit for USAID Food for Peace provisions distributed through the nonprofit organizations responsible for local schools in the form of free lunches. Third, instructor competencies were very weak and children did not learn a great deal. Schools were therefore seen as critical forms of child care and family food budget support much more than as academic or training institutions. And they were full to the brim.

These insights, which were the reverse of the official rationale for schooling but helped explain many things about its development in poor neighborhoods, grew out of careful attention to (and categorization of) what families said and did about their children's education and how they actually used their scarce resources (Easton & Fass 1989).

those already adopted. In fact, in qualitative data analysis categories and themes begin to have a life of their own, evoking new ideas and bringing new observations into focus.

Box 9.4 provides an example of a list of categories developed for research on an employment training program. Note that these categories include both descriptors of different aspects of the program and its environment and expressions, images or code words ("holdovers", "makin' a way out of no way at all") in the subjects' own language which flag important clusters of meaning.

*Box 9.4. Example of qualitative coding:
Codes from a study of job training programs (Berg 1995: 180)*

- | | |
|---|---|
| 1. Trainees' attendance | 19. Relationships among trainees |
| 2. Training center itself: appearance, reputation | 20. "Lying" |
| 3. Companies participating in the program | 21. "Dropouts" |
| 4. Staff definition of their involvement | 22. Counseling |
| 5. Staff views of trainees | 23. Referral meetings |
| 6. Company personnel managers' views of trainees | 24. Boredom |
| 7. Trainees' views of jobs | 25. "Killing time" |
| 8. Trainees' perspectives on work | 26. Poverty programs |
| 9. Trainees' views of staff | 27. "On-the-job training" |
| 10. Trainees' views of other trainees and self | 28. History of the program |
| 11. Recruitment of trainees | 29. Hustling |
| 12. Trainees' personal background | 30. "The cost of working" |
| 13. "Holdovers" | 31. Children |
| 14. Trips to factories | 32. Neighborhood living conditions |
| 15. Program success (how seen, assessed) | 33. Big business involvement |
| 16. "Hard core" | 34. "Counseling" |
| 17. Joking | 35. State employment service |
| 18. Follow-up | 36. Rapping |
| | 37. Money |
| | 38. The director |
| | 39. The Chamber of Commerce |
| | 40. "Makin' a way out of no way at all" |
-

BEST COPY AVAILABLE

Coding and grouping the data

Once one has a list of categories, classification of the data can begin. Each of the categories is assigned a number or a letter code. The data analyst then rereads or re-examines the entire body of material in order to identify elements (quotations, observations, pictures, particular artifacts) that seem to bear a relation to the different themes in the category list and to label them with these codes. Note that a single element - a quotation from some program participant, for example - may bear a relation to two or more categories on the list and will therefore be labeled with several codes.

The coded data can then be grouped by category for examination: for example, everything that was said by different parties and observed in the field regarding the competence of instructors can be considered together. There are now computer programs that facilitate this work considerably by providing an easy means of cataloguing and regrouping coded textual - and even graphical - data in a variety of ways.⁶ (Box 9.5 is excerpted from a qualitative researcher's own account of his data analysis experience. Note the researchers' annotations in the margins.)

Analyzing and understanding the data

As pointed out earlier, data analysis in qualitative evaluation is in large measure a question of "pattern recognition" - that is, of understanding what are the concepts and categories of importance to various stakeholder groups (which may mean office staff as much as it does community residents) and how they put them together, how they identify cause and effect, meaning and value in regards to the program under consideration. What do the concerned parties take to be the prime qualities of a good (or bad) instructor, and how do they see these being connected to different kinds of program results? What for them constitutes program "success" and how is it achieved?

These kinds of insights are derived from the qualitative data by carefully examining the material (quotes, observations, descriptions, artifacts) coded into the different categories in order to formulate - or, better yet, discover how stakeholders *themselves* express and formulate - the essential characteristics of each of these domains. How is motivation for training linked in their minds to gender, or material circumstances, or language of instruction? What real practical learning needs

Box 9.5. Example of qualitative fieldnotes
(Berg 1995: 178)

Vista City School Teacher's Lounge
February 3, 1991

use of
teacher
lounge

"they"

important
things

my own
thing

controls,
parent
conferen-
ces
responsibility

Then I went down to the teachers' lounge to see if anybody might happen to be there. I was in luck. JM sat at the first table, correcting papers; KT was also there, walking around and smoking. I said, "Hi! Okay if I join you?" "Sure," J said. "You and family have been to China, right?" I said, "Yes. Why?" J then turned to K and said, "Have you studied China yet? Sari has slides that she can show." K said to me that she was going to study world communities, even though "they" had taken them out of the sixth-grade social studies curriculum. "Now can you tell me who 'they' are?" I asked. "She said, 'You know: 'them', 'they'.'"

Both J and K were upset at how "they" had mandated what the teacher could teach in classroom. "They" turned out to be the central office [which] had communicated the state's revised sixth-grade social studies curriculum to the school. The state has "taken out all the things that we think are important" from the curriculum and has substituted the term of "economic geography" for the sixth graders to study.

Both J and K think that "sixth graders can't comprehend economic geography well" (J's words) and think world communities of Africa and Asia are more important. They said they planned to teach what they wanted anyway. K remarked, "They'll come around one of these days." "Oh, K, are you a rebel?" I asked. "No," was the reply, "I'm just doing my own thing."

After we chatted for a little while, J turned to me: "You're interested in what concerns us. I guess one thing is parents." She proceeded to describe a parent conference she had participated in yesterday afternoon with a child's parents and a child's psychiatrist. She said, "What really upsets me is how much responsibility they placed on me to change the child's behavior." They seemed to give lip service, J reported, to having "controls" come from the student. K's comment: "It's so difficult for parents to see that kids need to take responsibility [for] their actions."

do they see arising in the domains of health, agriculture and natural resource management and to what degree does the program speak to these? How do their viewpoints on these different topics diverge and what explains the differences of opinion? In a sense, the evaluator is reconstructing an implicit conversation among the principal stakeholders and trying to understand what its dynamics reveal about the underlying behavioral and attitudinal assumptions that affect the operation of the program.

Synthesizing the data

What one hopes to derive from this analysis is a good synthetic "picture" of how the program actually works at the local level, how it interacts with local institutions and culture, how key stakeholder groups experience the program in its setting and what sense they make of it. The "product" that results may take the form of a case study, a monograph, a story that evokes these dynamics, or more of a traditional analytic and expository presentation that probes the patterns of meaning and the local understandings of program operation that have been discovered. In virtually all cases, the presentation will make ample use of the actual words of those interviewed and salient elements of description of the phenomena observed in order to give the reader an immediate sense of their reality.

Validation

Systematic efforts must be made to verify with those interviewed or involved in the qualitative evaluation the accuracy and validity of what has been recorded and analyzed. In such "member checks", one may ask, "Is this a fair representation of what you said?" or "Am I accurately portraying what you mean?" To this extent, qualitative evaluation is always participatory. In a more fully participatory exercise, however, these "member checks" will be the occasion to invite those interviewed to reflect on the collective record of what was said and observed, to bring new experience or insights to bear upon it, and to participate in actually formulating the conclusions and/or final report. (See example in Box 9.6.)

A related technique is to attempt to formulate, on the basis of the insights about thought and meaning patterns that can be derived from the existing data, new relationships or understandings not broached in

*Box 9.6. Member checks in participatory research and evaluation:
The challenge of "restitution" in Senegal*

Qualitative researchers often speak of the importance of "member checks" - that is, efforts made after initial data compilation and analysis to verify with local informants that one has indeed "got it right", that their words have been fairly represented and that the relationships and meanings which the researcher has derived from observation make sense to them. In participatory evaluation, however, the notion of "member checks" takes on a new meaning, because local people are themselves to some degree co-authors of the evaluation.

However, though the subjects of participatory inquiry are very often authors or co-authors of the evaluation performed within or about their community, they are less often as thoroughly involved in the compilation, comparison and synthesis of results across entire regions or the country as a whole. Yet the role that they have played in evaluation at the local level typically equips them for - and gives them a live interest in - examining the broader results of the study, exchanging points of view with participants from other communities, and expressing their insights regarding overall conclusions to be drawn from the effort. Whence the importance of what is often referred to as the "restitution" stage in participatory evaluation: returning to the field, and/or to regional meeting places, with good synopses of overall results and some of the rich detail of the whole body of data in order to share these outcomes with those who took part at the ground level and, insofar as possible, to promote dialogue among them.

In the Republic of Senegal, a participatory research study conducted with the help of eight communities in 1995-1996 focused on the links between adult literacy and successful socio-economic decentralization. Village research personnel took part in all phases of the collection and analysis of local data, but the results were subsequently compared with those from similar studies in four other West African countries. The national research personnel involved took care, however, to bring back to the communities in question the preliminary results of this larger synthesis, in order to get their second-stage reflection on the outcomes of the study and the lessons to be learned from international comparison, and to facilitate at least a modicum of direct exchange with communities outside Senegal that had gone through the same experience. This feedback and exchange turned out to be nearly the most important part of the undertaking for the participants.

interviews; and then to verify with those concerned whether in fact this inference is correct. For example, though interviewees may not have talked at all of the role of women in cooperative management decisions, evidence from what men and women have said about the conduct of training courses, about women's current concerns and about the operations of the cooperative may lead evaluators to infer what is happening in this other realm. Checking whether this is the case will serve either to verify the accuracy of their derived understanding of the situation or to add new data and perspectives.

Since data collection and analysis are interwoven in qualitative evaluation, member checks often propel the evaluator to seek new data in order to further clarify patterns and understandings that are beginning to emerge. The rule of thumb in this regard is that data collection stops when one reaches the "saturation point" - when the material and insights brought in from the field begin to repeat and reconfirm each other and offer little new or divergent evidence. This will hopefully be because evaluators have succeeded in grasping a good part of what is happening and how stakeholders understand it, not because their interlocutors have learned to give them what they want or because they have managed to impose their own biases and blinders on the material to the point where nothing disconfirmatory is perceived!

An account of the qualitative data analysis process in a researcher's own words should offer a fitting conclusion to this section and serve to summarize many of the themes and methods just discussed. Erlandson (1994: 112-113) studied concepts of workmanship on the shop floor of an industry and describes how he assembled and analyzed his data in part as follows:

In developing this data base, I relied on the principle of constant comparison in which the most recent responses are compared with previous responses in the search for consistencies, discrepancies, anomalies, and negative cases. After each interview, I transcribed the interview verbatim and filed its material according to the categorization then in use. The material was typically in the form of paragraphs [that] were cross-classified to several categories. As I filed each statement, I compared it with previous statements in that category and kept running notes on the content of the category. The categories changed over time;

some disappeared and were merged under more general titles. Some emerged out of previous categories that became too heterogeneous. Some categories became parts of matched pairs or triads in which any given comment would typically be filed in each constituent category. For example, comments [that] described instances of lax work or bad workmanship also typically mentioned abusive management. Similarly, statements that described devising one's own procedures also typically included statements of satisfaction with the autonomy provided. This helped to reveal connections between categories.

I called this process to a halt when I felt that additional interviews were yielding little new information.

Putting it all together

The five steps in qualitative data analysis just reviewed (discovering themes and categories, coding and grouping data, analyzing and understanding this information, synthesizing the insights, validating the results) outline a method that is frequently used, though it is far from the only one. As mentioned earlier, researchers who put emphasis on the "artistic" side of qualitative inquiry would not necessarily follow the steps traced here but would proceed more intuitively and synthetically, cultivating a "connoisseur's" appreciation of the setting and the program. This amounts, in a sense, to proceeding directly to the fourth and fifth phases described above: synthesizing insights from one's overall perceptions and personal intuition, expressing these as artfully as possible, then validating them with stakeholders. In addition, the researcher or evaluator is often faced with the task of synthesizing multiple efforts of qualitative inquiry on related topics, some of which may have been conducted more in the artistic style, others in a more technical manner. Noblit (1988) presents a good introduction to this sort of "meta-analysis" across different studies.

Whatever the case, the evaluator need not necessarily go through the relatively full-blown procedures described in the previous section during every instance of qualitative inquiry. The common denominator of all such investigations is simply "thick description" of what is happening in the field and of how people perceive and interpret it.

Good qualitative inquiry demonstrates that the (sometimes invidious) distinction made in quantitative studies between descriptive and inferential levels of analysis is hardly a universally valid one. When you get down to it, it is not easy simply to "describe" things. To discern "what is really going on" amid all the phenomena and sense impressions that surround an educational program is no mean task. It requires a high measure of analysis, understanding, absorptive capacity and selective capacity. Putting chosen details together into an evocative and compelling picture of different aspects of the program is an art, and one of the greatest utility to the practical concerns of program improvement.

As previously suggested, qualitative investigation rarely stands full-fledged on its own in the evaluation field. It is usually coupled in an overall design with quantitative and statistical investigations. The quantitative data typically furnish an idea of the magnitude of the phenomena, whereas the qualitative data give a better and more suggestive picture of their nature and inter-relationship. The question, then, is how best to combine the two. Two separate orientations can be mentioned:

- In the first, qualitative inquiry serves to identify important themes and relationships in regard to a program, then quantitative inquiry is used to determine their incidence and magnitude and the likely degree of relationship among them.
- In the second pattern, quantitative investigation is used to explore apparent relationships among well-identified factors of importance to the program, and the qualitative investigation focuses on what these relationships and configurations actually mean to people and how they interpret and explain them.

Qualitative investigation is thus a critical element of program evaluation. After all, programs are essentially meaning patterns. If we do not know what they mean to people, and how they mesh with local customs, we understand very little about them. For this reason, qualitative evaluation should by no means be considered an exotic exercise to be carried out only by fully-trained and anthropologically certified personnel: this sort of practice in paying attention to what people are saying and doing at a local level is in fact a natural part of the training

of all program staff. In several programs with which I have been associated, special efforts were made to include representatives of line staff as field researchers in every exercise in qualitative evaluation. They turned out in most cases to be enthusiastic participants, and the experience left them with a number of new insights into the program and concerning the conduct of their own jobs.

Notes

¹ More detailed guidelines on the practice and process of qualitative research may be found in (e.g.) Bogdan & Biklen 1992; Denzin & Lincoln 1994; Glesne & Peshkiri 1992; LeCompte & Preissle 1993; and Morse 1994.

² In a number of Sahelian villages, the difficulties experienced with colonial administration led local residents to develop two parallel systems of governance: a bogus village chief and council for display to central authorities, and a real cast of local leaders who guided the community. Getting around one to the other (or even perceiving the difference) is in such a case no mean trick for outside evaluators. As much could be said of multinational corporations, military organizations and a raft of other social institutions that develop elaborate security systems to prevent unwanted access to operational information.

³ Obviously this judgment would have to be modified in cases where the language used by interviewees or local counterparts is itself the subject of evaluation or study, as might be the case if one were assessing the linguistic and curricular adequacy of instructional materials or trying to develop new ones. Under these circumstances, there would be a real reason for taking down all that was said and transcribing and analyzing it in detail, because the vocabulary would itself constitute important data.

⁴ In fact qualitative evaluation is closely allied in some ways to the whole tradition of participatory research, in which major emphasis is put on giving the supposed subjects or beneficiaries of social research a hand - and perhaps a determining voice in setting its objectives, developing its design, collecting the data and interpreting the results. See in this connection Whyte (1992) and Park (1995).

⁵ See in this regard the example presented in Box 9.5 in the data analysis section of the chapter.

⁶ "QualPro" software is one example of such tools. Tesch (1990) gives an overview of the use of software in qualitative data analysis.

References

- Bogdan, R. and S. Biklen. 1992. *Qualitative Research for Education: An Introduction to Theory and Methods*. 2nd ed. Boston: Allyn and Bacon.
- Denzin, N.K. and Y.S. Lincoln eds., 1994. *Handbook of Qualitative Research*, Thousand Oaks, CA: Sage Publications.
- Easton, P. and S. Fass (1988) "Monetary Consumption Benefits and the Demand for Primary Schooling in Haiti". *Comparative Education Review*. 33 (2): 176-193.
- Eisner, E.W. 1985. *The Art of Education Evaluation: A Personal View*. London: The Falmer Press.
- Erlandson, D.A. and Edward L. Harris, Barbara L. Skipper and Steven D. Allen. 1993. *Doing Naturalistic Inquiry: A Guide to Methods*. Newbury Park: Sage Publications.
- Glaser, B. and A. Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine Press.
- Glesne, C. and Alan Peshkiri 1992. *Becoming Qualitative Researchers: An Introduction*. White Plains, NY: Longman.
- LeCompte, M.D. and Judith Preissle. 1993. *Ethnography and Qualitative Design in Educational Research*. 2nd edition. San Diego: Academic Press, Inc.
- Morse, J.M. 1994. *Critical Issues in Qualitative Research Methods*. Thousand Oaks: Sage Publications.
- Noblit, G.W. 1988. *Meta-Ethnography: Synthesizing Qualitative Studies*. Newbury Park, CA: Sage Publications.
- Patton, M.Q. 1990. *Qualitative Evaluation and Research Methods*. Newbury Park, CA: Sage Publications.
- Punch, M. 1986. *The Politics and Ethics of Fieldwork*. Beverly Hills, CA: Sage Publications.
- Tesch, R. 1990. *Qualitative Research: Analysis Types and Software Tools*. New York: The Falmer Press.

CHAPTER 10

ADMINISTRATIVE EVALUATION

When I first became attuned to the importance of "administrative evaluation" about twenty years ago and first began personally promoting the idea in the nonformal education programs with which I then worked, the response was generally one of friendly incredulity. Why spend time evaluating the administration? Wasn't the point of any NFE program the training of beneficiaries in the field? Wasn't it vastly more important to gather information about what was or wasn't happening at that level?

My response was that what did or didn't happen in the field often bore a direct relationship to what had or hadn't been done by the organizations and administrations responsible for implementing these programs. In fact, after two years spent with local researchers in a painstaking participatory evaluation of one large-scale literacy program in West Africa, our main conclusion was that, whatever the deficiencies in course methodology and organization at the village level - and there were, of course, a number of these - the real obstacles to better results instead lay "upstream" and "downstream" of the teaching and learning activity *per se*. They lay in bottlenecks within the administrative apparatus responsible for the project on the one hand, and, on the other, in the lack of "post-literacy" opportunities in that particular rural environment.

A little bit of systems thinking, of the sort developed in Chapter Four of this book, makes the point quite evident. One can see quite clearly in Figure 4.3, for example, that it is the administrative organization responsible for teacher training, materials delivery and inspections that determines to a large extent the conditions in which nonformal education centers are created and in which they function thereafter. Even in a very decentralized system, central or regional authorities must still make sure that local authorities have the authorizations and technical support they need to do an effective job of promoting education and training. Administrative evaluation is no less important in this case. All that differs is that it is more focused on the regional or local level.

The trick is momentarily to consider the administrative domain of Figure 4.3 as a system unto itself (which it manifestly is from a number of points of view). Granted, the overall goal of those working in the

administration of an NFE program is to promote the learning and empowerment of intended beneficiaries in some specific ways defined by the programs' particular teaching, employment and/or community development objectives. That is presumably everyone's goal. But the question that is pertinent for administrative evaluation is much more concrete: within this general framework, what is the specific work that administrative personnel are expected to perform? What do they contribute toward accomplishment of the overall purposes of the program?

Defining administrative output

When one poses the question in this way it quickly becomes evident that program administrations all have one critical but intermediary function in NFE initiatives. Stated in the most general terms, it is *to produce and deliver to the field*, in the necessary quantities and at the appropriate times, *the specific complement of goods and services that local personnel need in order to carry out the program and ensure its success.*

I write "complement" because, of course, an important portion of the resources that such programs require originates directly at the local level. But often it is the outside contribution - whether of funds, skills, material, or political support - that makes the difference in catalyzing or "leveraging" local energies. This is obviously the situation with any project methodology "imported" from the central or national level, or transposed from other environments with the help of outside personnel, even though it may (and generally should) be adapted to local conditions before implementation. The idea in this sort of circumstance is either that the imported innovation will enable local actors to obtain results they have not previously been able to obtain on their own, or that the fact of participating in a program that spans a variety of localities will give them access to new resources and sources of inspiration. Even where the program is of strictly local origin, however, the fact that it is funded in part by extra-local institutions or hooked into a larger network generally means that outside support plays an important role in helping it to achieve its objectives and that the effectiveness with which those institutions play their role is worth examining. In any case, there is almost always some element of external infusion in an NFE program, and it is often a critical component in the overall "recipe."

It is therefore important to emphasize that the responsibility of administrative personnel lies essentially with the production and delivery of the material and services that make NFE programs possible (manuals, teacher training, inspections, etc.) - not with the actual field-level teaching or learning itself. In fact, though it is arguably essential for administrative staff periodically to "recycle" themselves through direct exposure to field conditions in a program, were this staff to all go to the field and begin teaching, the output of the program would probably decrease markedly - because no one would be minding the "tiller" or attending to those issues of policy, funding, materials development and teacher training that make educational programming possible in many more localities than the staff in question could ever frequent directly.

This, of course, is part of the fundamental rationale for a bureaucracy - namely, that by hierarchical deployment of personnel, it can spread the benefits of central knowledge and resource accumulation capacity in an equitable manner to far-flung localities and marshal their collective energies around a common goal. Bureaucracy also has its "down sides," of course, to which we will return in a moment. But it is more to the point right now to note that administrative personnel play precisely this sort of role in the "program theory" or "intervention model" of an NFE activity explained in Chapter Four. In other words, the overall "program model" generally proposes that certain services and materials provided centrally or, in any case, from the outside, be combined in specified ways with local resources in order to permit the implementation of an innovative set of activities, or the reinforcement and development of some set already at least latent in the local setting - and that these activities will have specific beneficial results for those most in need. In short, the specific "service and material production" function of the program's administration is an integral part of the innovation proposed and therefore merits evaluation every bit as much as the teaching and learning end of the program.

Of course, this administrative function can be broken down into a number of subfunctions (personnel recruitment and training, planning, accounting, materials development, etc.) corresponding to the different services and offices that fill out agency headquarters and that theoretically must work together in making sure services and material of the requisite quantity and quality get delivered in the field. Moreover, these functions are also distributed across levels of a program - e.g., central, regional, local - if it covers a broad area, and may be distributed in very

different fashions and carried out more or less effectively. And to perform the related work of program development and supervision, administrative offices obviously themselves require inputs - human, financial, material, conceptual and political resources that enable them to do their job and that they transform into goods and services to deliver to program participants. All these variations in administrative set-up, division of labor among and within central and branch offices, and coordination styles are part and parcel of the "methodology" of the organization in question. Figure 10.1 presents just such an enlarged portrayal in which the inputs, processes and outputs of the administrative domain itself are represented.

Accountability and "total quality"

Happily, this perspective on administrative evaluation no longer seems as strange and counter-intuitive as it did 20 years ago, thanks in large part to the movements of business restructuring and administrative reform that have been circling the world of late. Terms like "total quality management" (TQM), "site-based decision making" (SBDM), "management by the objectives" (MBO), "corporate re-engineering," "administrative accountability," and "quality circles" (QC) have entered the vocabulary of organizations throughout the world.¹ While there are variations in their meaning and in the strategies of organizational reform and renewal that they represent, the underlying notions might be summarized in the following seven points:

1. In an epoch of increasing democratization and accountability in the public sector and increasing free market competition in the private sector, organizations sink or swim on their ability to satisfy customers by producing quality goods and services in a reliable manner. "Quality" is defined as whatever best and most predictably meets the needs of customers, whether they are the consumers who buy the products of business firms or public citizens who seek the ministrations of government and non-profit organizations.
2. "Quality" is a moving target, since technology, customers' needs and the environment are all changing.
3. The reliable production of quality goods and services depends on the effective mobilization and participation of all organizational staff, and especially the "line workers" who actually put together and de

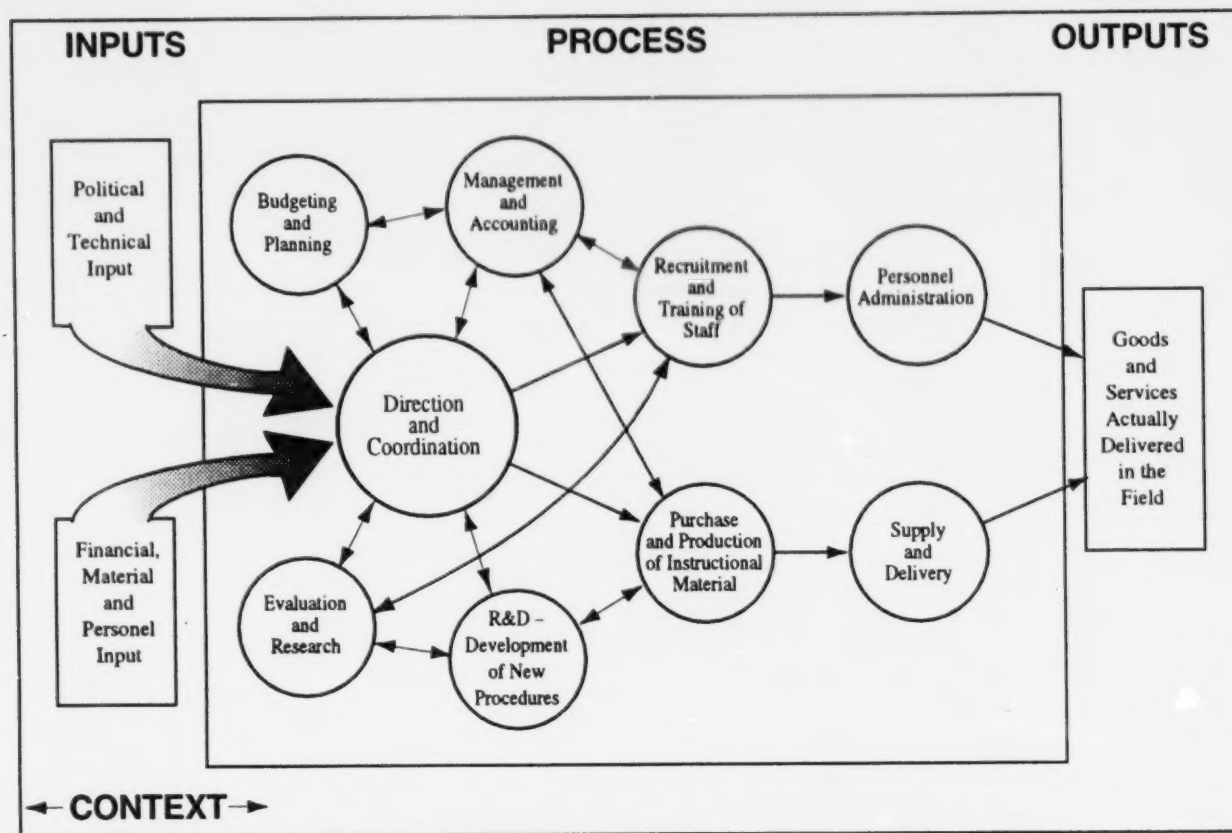


Fig. 10.1 Schematic representation of domain of administrative evaluation.

liver the products of the organization. In service industries and government agencies in particular, quality is greatly influenced if not determined by the capacities and activities of those people at the end of the chain (e.g. local staff or sales agents) who deliver services to the public.

4. It follows that problems in the production of quality goods and services, which inevitably arise from changing technologies and environmental conditions, can only be resolved by a team effort in which line personnel participate as equals with management and technical staff.
5. Organizations need therefore to restructure themselves to promote greater teamwork among personnel at different levels, to reward initiative and inventiveness in problem solution by such teams, and to instill a sense of shared responsibility for quality of production. To this end, TQM advocates insist that all workers have a sense of the overall operations of their organization and that units recognize the "internal customers" that they must satisfy in order to guarantee quality in the end product - that is, the other sections of the organization "downstream" from themselves in the chain or web of production to which they must furnish intermediate goods or services in order for the organization's final output to be achieved.²
6. In this sort of restructured organization, hierarchy is considerably flattened and supervisors must learn to function more as facilitators and instructors than as bosses.
7. Finally, in this sort of restructured organization, a premium is put on *individual and collective learning* to equip everyone for the new modes of team operation, the changing technologies of production, and the skills of creative problem-solving.

Why this suddenly-increased concern for the quality of services that businesses and administrations deliver? Analysts attribute the phenomenon to a number of different causes, according to their political stripe. For some, it is a natural consequence of an emerging world economy where many producers compete and both companies and administrations must develop their comparative advantage in meeting "customers" needs if they are to survive. Others point to the evolution of industry worldwide toward more of a tertiary or service sector orientation and observe (a) that services are mostly differentiated from each other by their quality, (b) that most of the quality of a service is determined at the

point of delivery, and (c) that improving quality under these conditions necessarily means giving more autonomy to lower-echelon workers at the point of delivery. Still others see in the TQM and administrative accountability movements a way for capitalists to rationalize downsizing their workforces and trimming government revenues in a time of falling profit margins: we give you less, but (purportedly) of better quality. Whatever the case, these current phenomena at least give evaluators a much-needed vocabulary for talking about the importance of administrative evaluation in promoting the effectiveness of educational programs!

Conducting an administrative evaluation

How then does one conduct an administrative evaluation? The first point to make is that this is *par excellence* an arena for participatory methods. All the reasons that justify participatory evaluation elsewhere in the program are at least as applicable at this level as well. In many respects, no one is better placed to figure out what isn't working in administrative procedures (and how to fix it) than administrative staff themselves, though - just as in the instructional or impact domains - participatory evaluation often works best if combined with some outside viewpoints as well. Moreover, administrative staff are much more likely to implement recommendations that they themselves have helped formulate than those drawn up by an outside "expert."

The TQM, restructuring and accountability movements furnish a range of examples of how to involve staff of an organization in the diagnosis and remedy of its operational problems. For example, the Xerox Corporation began by recruiting union representatives, line workers and management into problem-solving groups devoted to diagnosing why the corporation had lost market share in the competitive photocopy industry and - as their fortunes improved - gradually moved on to develop a whole training system and structure for quality improvement that they have also "exported" to other businesses and organizations (Pace and Argona 1991). In Box 10.1, another example of corporate restructuring illustrates the sort of administrative evaluation it has entailed.

Five elements in the conduct of an administrative evaluation will be briefly described below: building motivation, developing vision, analyz-

*Box 10.1. Administrative self-evaluation:
"Total Quality Management" at Johnsonville Foods*

Johnsonville Foods, a sausage manufacturing industry located in Sheboygan, Wisconsin, in the United States, adopted in the early 1980s a new style of management in an attempt to galvanize flagging productivity. Its owners decided to allow workers on the line to make all essential technical decisions about their own jobs and to provide them with the learning and training resources they felt they needed in order to improve productivity and, with it, their own earnings. In effect, teams of employees that cut across hierarchical levels and specialties within the company were invited to do their own evaluation of obstacles to improved company productivity, work out new solutions and implement them. The results have been very positive and have had a marked effect both on employee morale and on productivity.

The experience at Johnsonville Foods is similar to that in a number of companies and agencies from Europe to Japan and the Americas that have adopted, to varying degrees and with varying levels of success, what is currently referred to as the "Total Quality Management" (TQM) movement in industry and - more recently - government. TQM is an effort to increase quality of production and profitability through giving increased responsibility to employees for evaluating the organization and determining changes that need to be made in productive and administrative processes (Honald 1991).

ing administrative strategy, and enhancing team organization. Additional references and sources of information are listed at the end of the chapter.³

Building motivation

The first step, of course, is to convince staff that it is *worthwhile* to undertake this exercise - a problem that does not necessarily lessen when the exercise is mandated from above. What are the possible payoffs here for staff? One is the training component of any evaluation. Another is the opportunity to get a better grasp of the overall operations of the organization than most staff have ever had previously, and to get a more concrete sense of how each of them is contributing to the overall mission of the organization. Scott (1991) considers this a critical aspect of employee empowerment. A third is what one might call the

BEST COPY AVAILABLE

"convivial benefits" of working with colleagues from across the organization.

In addition, if the administrative evaluation is in some sense part of a larger participatory evaluation effort that involves field offices and beneficiaries as well, that overall configuration may add some other important elements of motivation. For one thing, administrative evaluation becomes an equity question: if local participants are evaluating their own performance, why shouldn't administrative personnel do the same? For another, the results of local participatory evaluations may pose, like a rising tide, a series of questions about the effectiveness of administrative functions that make assessment at this level imperative - if only to stem the discontent! There are therefore reasons for *timing* administrative evaluation so that it follows on the heels of participatory evaluation in the instructional and impact domains, because the results of these inquiries will certainly provide ample "grist for the mill" of administrative self-examination.

Reassessing objectives: What's our "vision"?

Once the motivational issues are handled in the best manner possible, the actual technical sequence of activities in administrative evaluation is very similar to that in any other branch of evaluation. An early concern will naturally be to reconstitute the strategy or "program model" that explains how the administration is supposed to function and how its various functions are supposed to converge in promoting effective service delivery in the field: in short, the configuration portrayed highly schematically in Figure 10.1. But that is not necessarily the first step. Remember that in some sense the administrative apparatus of a program - whether it is within a public agency, a private enterprise, a nonprofit organization or a community association - is the entity responsible for setting its goals, designing its strategy, and providing guidance to those working in the field. The commitment of administrative staff to the program will depend in large part on the degree to which they identify with its goals and objectives and feel both aware of and personally engaged by the vision of the future that it offers: the future of the organization itself and the future for the beneficiaries that it serves. Yet the goals and objectives of an organization are not always clear and well defined; they are frequently handed down from legislation or previous administrations; and they

have a tendency to age quickly as circumstances change - both the circumstances that the program is designed to address and those in which its administration must operate.

It follows that the "vision" of most programs requires regular revisiting and renewing. This is a sort of evaluation that may take place prior to assessment of the actual administrative performance of the organization - or it may be triggered by the results of such an assessment; and it is an exercise that benefits greatly from broad-based participation. Box 10.2 presents an example of a "visioning" exercise in which contextual data and performance analysis were paired with the evaluation of goals and objectives *per se* and in which broad

*Box 10.2. Evaluating organizational objectives:
"Visioning" in a graduate School of Education*

"Visioning" is a methodology for re-examining the sense of mission of an organization and the image its members have of the future that they would like to create for (or through) it. Methods for visioning have become increasingly popular in both the public and private sectors of a number of countries as a means of institutional renewal. Visioning amounts, in effect, to re-evaluating the objectives of the organization: not just have we met our goals, but do we still think they are worth meeting? And what goals would we really like to accomplish?

An activity of this sort was carried out in a College of Education (teacher training institution and graduate school for childhood and adult educators) of the State of Florida, USA, over academic year 1995-1996 - with a difference. Whereas visioning exercises are often confided in a small management committee or a task force with some representation from other levels within the organization, the Florida team sought both to associate all levels of personnel within the College of Education (faculty, students, staff) through a representative-constituent model, and to hold hearings with representatives of what it identified as the principal "stakeholder groups" of the College: elementary and secondary school teachers and parents, professional associations, the business community, state and local policymakers, etc. "Vision" was interpreted in this case not just as a group's wish list for the future, but as a product of the interaction between the values of College personnel and stakeholders on one hand, and their sense of the challenges arising from environmental change and institutional performance on the other. Evaluators played a significant role in this process.

BEST COPY AVAILABLE

participation was sought from the outset. Walzer (1996) and Weisbord (1992) offer further detail on visioning methods now widely used. Evaluation is, in fact, a natural concomitant of visioning: if you don't know where, it's harder to make sense of where you are going!

Analyzing the administrative model

Sooner or later, therefore, renewal of an organization's vision adds little of value if the administration cannot also get a "fix" on its own performance and assess the success with which - and the way in which - it is meeting its own operational objectives as the unit responsible for supervision and support of the program. To make this assessment, one must be able to paint some kind of picture of how the organization is intended to operate and how it is in fact operating: that is, compare its model with its daily operations.

One way to tackle this task is to start by asking just what these services and materials *are*, to whom they are supposed to be delivered and when, in what quantity, of what quality, and for what reason - i.e., what role these inputs to the instructional domain (which are in fact the "outputs" of the administrative one) play in making the program more effective and productive. Once these intended outcomes of administrative activity are defined, one can begin to describe how the work necessary to produce and orient them is divided up and accomplished: in short, the "methodology" of the administrative organization. And this leads back in turn to the question of the resources (human, material, financial) which the organization needs to function, and the way in which it uses them and accounts for them.

Obviously as one progresses through this inquiry, the "model" of administrative activity quickly ramifies and one discerns a whole series of subfunctions and connected units, each of which depends on others within the same administrative framework to accomplish its particular work. In the TQM literature, as mentioned earlier, a distinction is often made between "internal customers" and "external customers." The latter are the people to whom the organization finally sells or distributes its services or products. But in order to do so it must organize its own internal chain of production in such a way that certain offices or units prepare material and services for others, and so on down the line or throughout the entire web of the organization. If the curriculum development unit never puts out decent material for administrator or teacher

training, the personnel unit can hardly be expected to have much success supplying people with the necessary competence to the field. The curriculum unit must therefore cater to an "internal customer" - the personnel office (for whom, moreover, the curriculum specialists are "internal suppliers") - before either of them can do much for their external customers.

Figure 10.1 above illustrates these relationships quite schematically. The larger "context" does not really appear: it is constituted essentially by the entire program for which the administrative units work, as well as by the general environment in which both the program and its administration operate. As for the "impact" or longer term outcomes of administrative behavior, they are in a sense identical to those of the program as a whole, though innovations adopted in its organization may also have a lasting effect on administrative culture in the general social system. (These would probably be "unintended" effects of the program). But the general production process of the administrative domain can be clearly distinguished in the diagram. To the degree it does not work, someone or something else must assume these functions, because programs cannot operate for long without some source of needed inputs.

Team organization

Once one has begun to develop a sense of the operations of the entire administrative apparatus, a second evaluative question quickly arises: what are key indicators of operational quality within the different components of the organization? What three or four (or twenty) types of data best portray the quantity, quality and timeliness of service or material delivery in the field? And, as one moves back up the "map" of administrative strategy, which indicators of the effectiveness of the different internal suppliers and consumers would give everyone the clearest idea of what is and is not working, and of what could be done to improve the productivity of the entire organization? These types of data then constitute the indicator matrix for administrative evaluation.

Each unit within the organization should have a hand in choosing and/or devising its own indicators and in deciding how to use such data. That approach is implicit in the notion of participatory evaluation. In fact, the simple exercise of beginning to detail and analyze the workings of an administrative organization in this way may already provoke or

liberate some innovative thinking on the part of staff. It often becomes evident that they know what most of the problems are already and can identify many of them and envisage solutions before the first bit of performance data has been recorded or examined. It is nonetheless useful to forge ahead and progressively put in place mechanisms by which staff can monitor their own productivity. But the real point of the exercise is precisely to stimulate the innovative and creative response of personnel to organizational problems and thereby find new ways to support the program in the field.

It would seem natural to subdivide the administrative evaluation by unit or productive grouping within the organization - once this map has been established or reconfirmed to everyone's satisfaction - and then work upwards from this piecemeal evaluation to an assessment of the overall operation of the organization. But there are also benefits to assembling "cross-functional" teams that include people from several different branches and levels of the organization in order to study particular problems of work flow and productivity rather than the single activities of individual offices. This is especially true wherever the internal division of labor in an organization has become "calcified" or rigidly entrenched in a particular hierarchical pattern of units and subunits which no longer serves much productive purpose or even impedes work. A participatory administrative evaluation of this nature may well serve as the starting point for reconfiguring and simplifying the internal workings of the organization.

In a large organization or agency, administrative evaluation may in fact be a several-tiered undertaking, since there are administrative units at all levels: center, regions and (often) localities as well. One needs in this case to pay particular attention to the division of labor and flow of work *between* levels of the organization; and the notion of "cross-functional teams" here obviously signifies working groups that include a "vertical cross-section" of central, regional and local staff as well as a lateral one at any given level.

Nurturing a professional bureaucracy

In a paper on the evaluation of program implementation delivered at the UIE/DSE workshops on evaluation methodology, Alain Dunberry calls attention to the relevance for evaluators of Henry Mintzberg's classic

distinction between a mechanistic bureaucracy and a professional bureaucracy (Dunberry 1993). The latter is much "flatter" in form: that is, much more weight is given to the professional work of its field staff and the hierarchical distance between local and central agents is much less than in a traditional bureaucracy. Because the "center of gravity" of the organization, and the preponderance of its technical competence, is much closer to the field, it is able to adapt to complex and varied environments much more easily than a traditional bureaucracy, which tends to want to impose standard solutions everywhere, in a cookie-cutter fashion, by virtue of its "authority." In short, a professional bureaucracy is no less coordinated, but much more decentralized.

Arguably participatory administrative evaluation has a tendency to transform traditional bureaucracies into professional ones and to promote a certain degree of flattening of hierarchy. Insofar as administrative staff at every level become more acutely aware, through this evaluative process, of the overall operations of the organization and of the factors that promote or impede its productivity, they also become increasingly capable of making decisions about their own work without micro-management from above.

Much of the data for administrative evaluation is of a descriptive quantitative nature - what sort and quantity of services were rendered when and where, and how does this compare with norms, expectations and/or needs? (In the business world, this sort of assessment usually entails comparison with the performance of competitors or industry leaders and is termed "benchmarking.") But there are important qualitative elements as well. For one thing, the levels of satisfaction of "internal and external customers" regarding the quality of services rendered, and these customers' analyses of problems that have occurred will be a critical barometer of success. In public agencies in North America, "customer satisfaction surveys" are currently very much in fashion.

For another, many factors that govern the productivity and harmony of a work environment are difficult if not impossible to capture with quantitative indicators of operations, or even with various satisfaction surveys. Each organization has its own internal "culture," replete with special ritual terms, interaction patterns and behavioral norms. It can be most interesting to invite staff or evaluators from one branch of the organization to spend time working in another - or in a different organization - simply to describe "ethnographically" the way in which they

experience such local norms. (Schwartzman [1993] provides a whole methodology for doing ethnography in organizations.) In any case, the current push for democracy and accountability within businesses and agencies requires the members of these organizations to become much more aware of how they function and why they malfunction. The kind of insight that qualitative evaluation can provide into the actual operation of an institution and the motivations which drive its personnel may be most useful in the effort of productivity improvement.

Conclusions

Where does administrative assessment fit within the overall design of an evaluation? All my own experience indicates that it is a critical but often neglected part. Three factors in particular underline its great importance.

First of all, it is generally difficult to analyze or explain the results of an NFE program without taking account of the support that it has or has not received from the organizations - be they local, regional, central or some combination of all these - responsible for conducting it. And it is even more difficult to identify how the situation might be improved without understanding the operations of these sponsoring units, which are generally responsible for supplying a good portion of the material and services needed for instructional work in the field, and without examining how their role might be more effectively played. "Organizational learning," to use the term first popularized by Argyris and Schon (1978), seems to be a prerequisite to improving the effectiveness of NFE programs.

A second reason for giving greater emphasis to administrative evaluation derives from a principle of equity and fairness. "What's good for the goose is good for the gander," as the saying goes. If the performance of field personnel and program beneficiaries is to be scrutinized, so then should be that of headquarters and central agency personnel. Current movements for increased decentralization and accountability only make this principle still more valid.

A final reason often comes to the fore in participatory evaluation. Most participatory evaluations involve giving local program beneficiaries greater voice in judgments and decisions made about the program. The reaction of middle echelon and field staff of NFE organizations in such a case, however, is not infrequently, "Why should we sweat and

strain to give local people greater voice in these matters if we ourselves have never had anything much to say in the evaluation and planning of our own agency?" In short, at some point for evaluations to be truly participatory at the local level they have to become more participatory at the intermediate ones as well. Agency staff need opportunities to play an enhanced role in evaluating their own sphere of activity if we wish them validly to assist local people in doing the same; and the sphere of activity and natural subject of evaluation for these staff is the domain of administrative operations.

Because the purpose of administrations is to provide services and material to those charged with implementing a program in the field, administrative evaluation often logically takes place once the results of the assessment of field programs come in. Part of this field assessment inevitably concerns the adequacy of support received from institutional sponsors and supervisory agencies, and that offers a natural point of departure for the administrative self-examination. In fact, it may be a very useful and productive practice to associate representatives of beneficiaries and "customers" in this brand of evaluation - in part for the insight and impetus that they can give to analysis of the performance of sponsoring organizations, and in part for the positive effects of the imagery itself: reversing the tables on normal hierarchical relations and encouraging those at the bottom of the chain to evaluate the performance of those at the top. In business organizations, this is often referred to as "upside-down personnel assessment" and is more and more commonly practiced in North America (cf. Petrini 1991).

Much obviously remains to be done and developed in the domain of administrative evaluation within NFE and training programs, as it does within formal education systems as well. This chapter has only served to open up the topic and suggest a few directions. But it is a fair bet that the subject will take on increasing importance in the years to come, as a means of "reinventing government" - to use the phrase in current vogue in North America - and of ensuring increased transparency, accountability and participation in the operations of other public and private organizations.⁴ The era of passive constituencies seems to have past, and administrative personnel have an interest in evaluating their own services before others do so!

Notes

¹ For examples of TQM literature in the private sector, see Bounds 1995 and Maromonte 1996; and, in the public sector, United States General Accounting Office 1995.

² In this sense, program participants at the local level represent the "internal customers" of administrative staff in their joint effort to accomplish program objectives.

³ Further useful material on the conduct of administrative evaluations may be found in Burnham (1995) and Love (1991).

⁴ Osborne (1992) is the original text on "reinventing government," propelled to prominence when it was frequently cited by US President Bill Clinton. There have been a number of sequels.

References

- Argyris, C. and Donald Schon. 1978. *Organizational Learning: A Theory of Action Perspective*. Reading, MA: Addison-Wesley Publishing.
- Bounds, G.M. 1995. *Management: A Total Quality Perspective*. Cincinnati, OH: South-Western College Publishing.
- Burnham, B.B. 1995. *Evaluating Human Resources, Programs, and Organizations*. Malabar, FL: Krieger Publishing Company.
- Dunberry, A. 1993. "Evaluating Organizational Performance." Paper given to the UIE/DSE Seminar on Evaluation of Nonformal Education, Berlin, June 1993.
- Honald, A. 1991. "Empowerment at Johnsonville Foods." *Training* 27(2): 18-23.
- Love, A.J. 1991. *Internal Evaluation: Building Organizations from Within*. Newbury Park, CA: Sage Publications.
- Maromonte, K.R. 1996. *Building the Invisible Corporation: The Executive's Guide to Transcending TQM*. Westport, CN: Quorum Books.

- Osborne, D.E. 1992. *Reinventing Government: How the Entrepreneurial Spirit is Transforming the Public Sector*. Redding, MA: Addison-Wesley Publishing Company.
- Pace, L.A. and Dominick R. Argona. 1991. "Participatory Action Research: The View from Xerox." In: William Foote Whyte eds., *Participatory Action Research*. Newbury Park, CA: Sage Publications.
- Petrini, Catherine.- 1991 "Upside-down Performance Appraisals" *Training and Development*. July: 15-22.
- Schwartzman, H.B. 1993. *Ethnography in Organizations*. Newbury Park, CA: Sage Publications.
- Scott, C.D. 1991. *Empowerment: Building a Committed Workforce*. Los Altos, CA: Crisp Publications.
- United States General Accounting Office. 1995. *Federal Quality Management: Strategies for Involving Employees*. Washington, DC: Author.
- Walzer, N. 1996. *Community Strategic Visioning Programs*. Westport, CN: Praeger.
- Weisbord, M.R. (Ed.) 1992. *Discovering Common Ground*. San Francisco: Berret-Koehler.

CHAPTER 11

SELF-DIRECTED LEARNER ASSESSMENT

Literacy campaigns and nonformal education programs are often not very hospitable environments for traditional examination and evaluation techniques. Approaches that are reminiscent of scholastic tests or administrative audits are likely in these circumstances to scare away participants or, at the very least, distort the data obtained and perhaps bias the results of performance measures. Evaluators have enough difficulty in obtaining a good living portrait of a program and its effects without bringing other negative influences to bear on the validity of their information. Moreover, since the purpose of most of these programs is, in one way or another, to "empower" participants and give them a sense of efficacy and self-direction, any evaluation technique that seems to relegate them again to the status of inferior, subject or pupil is *a priori* suspect.

Earlier in this book I came out strongly in favor of participatory evaluation as an approach well adapted to the needs and potentials of nonformal education programs, and technically just as challenging and satisfying as any more traditional approach evaluators might take to their work. It is time to suggest in a bit greater detail of what such a participatory methodology might be made. Most of the methods and examples cited in this chapter are drawn from over twenty years of work with participatory evaluation methodologies in literacy programs of West and North Africa and the Caribbean; but the basics involved should be much more widely applicable.

When we speak of "self-directed learner assessment," one thinks automatically of the evaluation of learning outcomes (knowledge and skill acquisition, attitude changes) since these are generally the heart of an educational program. But as systems analysis make clear, such results are only part of the overall process of nonformal education, and outcomes at this level cannot be understood without taking into account what is happening in other realms of the program as well. In the methodologies presented here, learner assessment of instructional results is therefore set within a framework of overall participatory evaluation of program inputs, processes, outputs and outcomes.

An example of participatory procedure

The prime example with which I am familiar comes from a methodology for participatory evaluation of functional literacy programs first developed and used in the Republic of Mali in the late 1970s but replicated, modified and improved in a number of other contexts since then. The program under evaluation in the initial case - a training program linked to a large-scale regional agricultural production and marketing operation - covered several thousand villages in western and central Mali and had been in existence for six years (Republic of Mali 1978; Easton 1983).

The basic notion of the participatory evaluation was that each village, neighborhood or organization responsible for a literacy center would draw up a "balance sheet" or "report card" its activities in order to serve as a basis for its own subsequent planning as well as to share with other centers and with regional headquarters. That report card had three parts, corresponding (the reader will quickly recognize) to the three domains of the systems model proposed in Chapter Four: administrative operations, or all the activities necessary to set up the literacy center and keep it running; instructional activities, or all the actual teaching and learning that went on during and around the NFE program itself; and social impacts, or all the ways in which the knowledge, skills and attitudes acquired in the program were applied thereafter by those who had benefited from the training and the effects generated in that way.

Within each of these domains, moreover, one could talk about resources and materials used ("inputs"), work methods adopted ("processes") and results attained ("outputs") - though it was not necessarily essential for everyone to go to this level of detail. The overall configuration of the report card or balance sheet could therefore be roughly pictured in the manner already indicated in Table 4.1

The participatory evaluation method then consisted, first, of determining the results of the program - classes held, attendance rates, levels of knowledge actually acquired, uses to which this knowledge had been put, and numbers and categories of people affected in each case - and then analyzing and discussing this operational balance sheet. Was it satisfactory, given the resources available or invested, in the opinion of participants? What explained the good, mediocre or poor results observed? Were the resources and methods used in these different domains of literacy programming adequate to get the kind of results

everyone was hoping for? If not, what most needed changing and how might the changes be effected? This sort of "three-dimensional" report card thus lent itself not just to description of outcomes, but also to a *diagnosis* of the situation.

How, though, could "results" be determined and measured? As far as the regularity of courses and of attendance was concerned, recording and examining the data was obviously not too difficult, providing that the basic information had been written down in one fashion or another over the course of the year. The evaluators needed only to furnish a form which simplified the compilation and synthesis of information from these existing sources. Learning outcomes and actual applications or impacts of this knowledge were a different matter, however. In most programs where participatory self-assessment methods were first applied, there already existed a welter of examination methods for evaluation of literacy outcomes, and very little if anything for assessment of post-literacy outcomes. Neither of these situations was satisfactory: for the most part, the tests used for gauging learning outcomes were normative-style examinations graded on a scale of 10, 20 or 100 where there was little evident equivalence between any given score and a definite level of skill or knowledge; and no one could say much about applicability of this knowledge. In the following two sections, we examine each of these problems: learner assessment of literacy results, and participatory assessment of post-literacy or post-training impacts - and the solutions devised for them in the Mali case.

Learner assessment of literacy results

Whole dissertations and numerous books have been written about proper methods for defining and measuring the acquisition of literacy. In the United States, two major new instruments have come out in the last few years: the National Adult Literacy Survey (NALS) and the revised Test of Adult Basic Education (TABE 7 and 8). The former requires about 45 minutes of individual testing and interviewing and yields an assessment of the adult's degree of mastery of a whole series of conceptual and practical skills related to the North American definition of "literacy."

This is obviously an area in which much research remains to be done and where a great deal of refinement of approach is possible.¹ But a distinction should be made here between research and evaluation. In the

latter, high levels of refinement are not always necessary. The key question for assessment is "How much accuracy do we need and what underlying purposes are we trying to fulfill?" In the Bayesian statistical approach to test construction, for example (e.g. Pollard 1986), it is acknowledged that no test is ever 100% reliable or valid and that the criteria to apply in test construction depend on the decisions that we want to make on the basis of the results as well as on the probability of accuracy that we require.

When the question is posed in this manner, evaluators realize intuitively that there is often a trade-off between refinement of approach and breadth of assessment. What can be done by a team of researchers with a sample of learners cannot easily be accomplished if one is concerned to determine results across an entire program. (A good sample survey will give relatively generalizable information about program-wide results, it is true, but will not furnish information that each site and locality can use in improving its own practice or in assessing the strengths and weaknesses of student learning.) There may also be a trade-off between sophistication of results and degree of involvement of learners themselves in their determination. If the objective of the assessment procedure is both to measure results with a usable level of accuracy *and* to serve as a stimulus to the program and to the learners themselves - as is typically the case with participatory evaluation - then a different method may be called for. To help us define what kind of method, let's consider for a moment the two parts of the question cited above (appropriate level of accuracy and underlying purpose of assessment) and the way in which this question is typically answered in participatory evaluation.

Appropriate level of accuracy

As the example of Bayesian statistics makes abundantly clear, the basic accuracy question concerns just what information one needs to get from assessment and how it is to be used. In many cases in NFE programs, staff or central offices know relatively little about the actual level and type of knowledge acquisition of participants. To get a first approximation - e.g., simply to determine in a regular and relatively reliable manner how many learners by locality or by category of persons involved have become literate (numerate, skilled) or semi-literate (semi-numerate, semi-skilled) - might already constitute a vast and significant

improvement on present levels of information. There is a French saying that "the best is a sworn enemy of the good" (*le mieux est l'ennemi du bien*): in other words, attempts to obtain highly precise and refined information may result in getting no usable information at all, or too little data to serve management purposes.

Underlying purpose of assessment

In most NFE programs, the underlying purpose of learning assessment is in fact twofold. On the one hand, program staff wish to get a sense of the actual capabilities that people have acquired through their study; and, on the other, they wish to motivate teachers and learners to assess their own level of accomplishment and to draw lessons from their successes and mistakes.

Insofar as learning assessment is practiced at all at local NFE sites, it is often a variety of assessment that accomplishes neither of these purposes. Most existing testing and assessment procedures have been copied from primary or secondary schooling and involve administration of examinations and grading of the results on some standard scale. This approach typically has two weaknesses:

1. It is not at all clear what real competence any particular performance represents. In other words, one does not really know just what a score of 12/20 or 85/100 means in practical terms or what the person who obtained it is actually capable of doing. School tests are typically "norm-referenced" ones - that is, they are tests designed to rank students with respect to each other for purposes of deciding who should be promoted or admitted to the next level, not to say exactly what competencies they have acquired.
2. The examination procedure itself is often a rather alienating one, particularly for adults who may have bad memories of previous schooling experiences or fears regarding how they measure up in an "academic" setting.

Realization of these facts has generally pushed NFE evaluators toward the conception and use of "criterion-referenced" tests - that is, exercises designed not to rank learners with respect to each other, but rather to determine just what they are capable of doing.² Can this participant decipher a few words, read sentences of intermediate

difficulty, read a short text with good comprehension, or handle a longer and more complex text? Can the person in question read numbers, perform simple additions and subtractions, perform multiplication with carryover, handle long divisions, or even resolve mathematical problems requiring the use of several operations? Criteria may also be devised that are directly referenced to specific practical tasks, like filling out an employment application, balancing a checkbook or materials inventory, reading warning labels on medicine bottles, or repairing irrigation equipment.

There is another advantage to such an approach: it is that these criteria have a good deal more meaning for learners themselves than does an abstract number grade. The learners can consequently be more actively involved in the testing procedure, at least at the level of interpreting and using the results. In a participatory learner assessment, however, one typically tries to go further still in the direction of enabling the learners to determine their own levels of knowledge acquisition.

Self-evaluation instruments

In the African cases of learner self-assessment, evaluators defined a scale with somewhere between three and eight levels for gauging the degree of literacy, numeracy or technical competence of learners. Sample problems were then generated for each level in question and displayed in a simple yet attractive "self-evaluation brochure." The brochures were distributed widely well in advance of the actual evaluation date and learners were invited to assess their own level of competence on the different scales: i.e., "I can do this much, but I haven't yet mastered that level." (Excerpts from self evaluation brochures used with literacy program participants in Burkina Faso and Morocco in recent years appear in Figure 11.1 and 11.2.) On the day of the assessment, each learner was asked to report on the results of his or her own self-evaluation. Monitors, teachers or other local personnel then congratulated each participant on his or her work and proposed to them another problem of the same level of difficulty drawn from an assessment manual likewise prepared by the evaluation team and disseminated to all concerned staff. This cross-validation allowed the staff person to "adjust" or correct the self-assessment of the learner where necessary. Results were then inscribed in the learner's own brochure and on the literacy center's "report card" or "balance sheet." Once all "scores" were recorded in this fashion, the information served

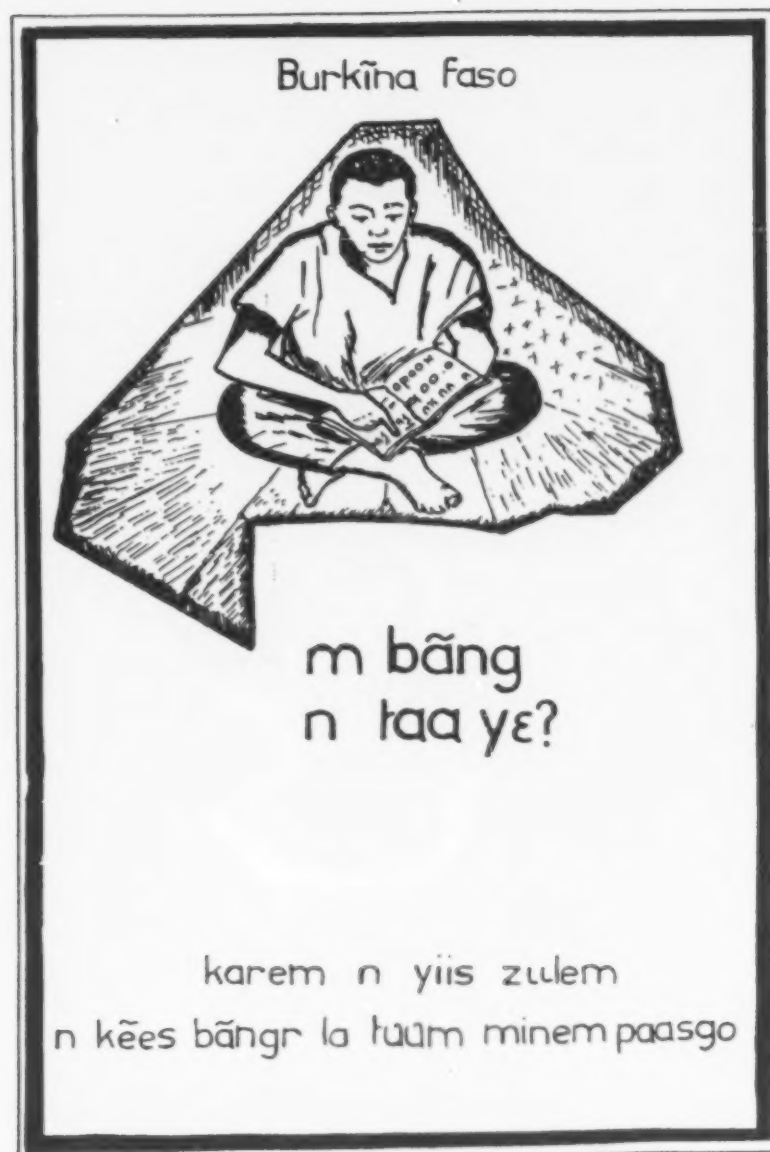


Figure 11.1a. Self-evaluation booklet from Burkina Faso (cover)

kaorengo		kaorengo					
1	<table border="1"> <tr> <td>a</td> <td>o i a ũ aa ā uu</td> </tr> <tr> <td>b</td> <td>ko ma do bu di</td> </tr> </table>	a	o i a ũ aa ā uu	b	ko ma do bu di	4	wakat bāg la tōnd kambā sungd yaalga? tōnd zae n yāame lē naoor wūsgo, biig sā n doge, a kuug a ye n lē tōag a kiis a yoobā lāa, a yaa sōma, n nob n tar yamleogga. kiis a yoobā suka, a ma wā lara blisem hal wūsug lē sekd biigā: sā n wa loag kiis a yoob lēka, tōnd yāame l' a vima ye ka noom a ye. a ma wā blisem boogdame. wakat kāng me, biigā sēn bulā yīnga, a rubā me raab paasdame. a sēn yaa bilā me yīnga, a pa tōe n dū kāsembā rubā ye. rē n kuld lē wakat ninga, biigā mi n faage, lē bāasā pa basd a lēle ye. rē marsā, yaa modgr lē d maan rub laare, sēn yaa biigā rub bala, n kāl a. yaa sud lē biig sā n tar kiis a yaobe, a pa tōe n dū rub ning yūlm pistā ned sēn dūl ye.
a	o i a ũ aa ā uu						
b	ko ma do bu di						
2	<table border="1"> <tr> <td>a</td> <td>buudu boko kundi kuga puugo</td> </tr> <tr> <td>b</td> <td>goama zag-neere ma-biiri nanguri</td> </tr> </table>	a	buudu boko kundi kuga puugo	b	goama zag-neere ma-biiri nanguri		
a	buudu boko kundi kuga puugo						
b	goama zag-neere ma-biiri nanguri						
3	<table border="1"> <tr> <td>a</td> <td>ned fāa rat a lo sāngre . biiga rabda raaga .</td> </tr> <tr> <td>b</td> <td>ziig sa n lik paalem, kood ligd ka toag ye d segenda ziig n ning d rūmsā. tōnd sā n bas d rūmsā lē b gōod yaare, ad yaa bōne . sēoaga, b na n kēe neb pul n wāb ki, bēnga. b sā n kē nina sūmkaang bi a goanga, yaa sāoonga.</td> </tr> </table>	a	ned fāa rat a lo sāngre . biiga rabda raaga .	b	ziig sa n lik paalem, kood ligd ka toag ye d segenda ziig n ning d rūmsā. tōnd sā n bas d rūmsā lē b gōod yaare, ad yaa bōne . sēoaga, b na n kēe neb pul n wāb ki, bēnga. b sā n kē nina sūmkaang bi a goanga, yaa sāoonga.		
a	ned fāa rat a lo sāngre . biiga rabda raaga .						
b	ziig sa n lik paalem, kood ligd ka toag ye d segenda ziig n ning d rūmsā. tōnd sā n bas d rūmsā lē b gōod yaare, ad yaa bōne . sēoaga, b na n kēe neb pul n wāb ki, bēnga. b sā n kē nina sūmkaang bi a goanga, yaa sāoonga.						

Figure 11.1b. Self-evaluation booklet from Burkina Faso (reading test in Moré)

- 1 -

A القراءة

أذكر الحُرُوفَ التَّوَجُّودَ مِنْ بَيْنِ العَلَامَاتِ الَّاتِيَةِ :

←	ع	س	و
ل	» «	:	م

10 □

اقرأ الكلمات الآتية :

عَمُّ	حَقْلٌ	طِفْلٌ	مَدْرَسَةٌ
طَبِيبٌ	مَسْجِدٌ	تَلْقِيحٌ	مُسْتَشْفَى

20 □

اقرأ الجُمْلَ الآتِيَةَ :

كَتَبَ عَزِيزٌ رِسَالَةً
زَيْنَبُ وَمَرْيَمُ مُشَارِكَتَانِ فِي النَّادِي السِّيَرِيِّ

30 □

اقرأ النَّصَّ الْآتِي :

تَرَكَ الرَّايِي أَغْنَامَهُ تَزَعَى فِي الْعَابَةِ ، وَجَلَسَ تَحْتَ شَجَرَةٍ ، فَعَلَبَهُ النَّعَاسُ . أَبْقَطَهُ نَبَاحُ كَلْبِهِ ، فَوَجَدَ الذِّئْبَ يَتَصَارَعُ مَعَهُ .
--

40 □

Figure 11.2. Self-evaluation booklet from Morocco (initial portions of reading test in Arabic)

BEST COPY AVAILABLE

as one basis for collective assessment of the results of the program. This approach has proven to have both a number of strengths and some weaknesses that can generally be mitigated by careful design. The strengths include -

- the active involvement of learners in determining - and reflecting on - their levels of knowledge acquisition: participants are typically strongly motivated by the experience, which in addition gives them a concrete sense of how far along they have gotten and what remains to be learned. Test anxiety can be largely overcome by the practice of distributing the test booklet well in advance of the assessment session and explaining the self-evaluative method.
- the quick turnaround that this sort of assessment makes possible: results can be determined on the same day and thus serve as a basis for broader local evaluation;
- the usability and interpretability of results at regional and central levels: the data reported make it immediately evident who can do what;
- ease of use and understanding by local staff, who can be trained relatively rapidly in the methodology and can pass the training on in "cascade" fashion: it is entirely possible, with this approach, to associate local staff in test design by enlisting their help in validating the definition of levels and elaborating the supplementary problems that will go into the evaluation manual; and, finally,
- the training in the mechanisms of evaluation that the experience affords local staff and learners alike.

In fact, that experience often proved to be generalizable in unexpected ways: NFE participants who refine in this manner their ability to draw up a "report card" or "balance sheet" on a collective undertaking like adult literacy may then decide to apply the same type of "technology" to assessing the results or productivity of local agricultural cooperatives, reforestation programs, and public health campaigns. (See the example in Box 11.1.) This is no accident. Since their invention on the irrigation schemes of the fertile crescent several thousand years ago, literacy and the technology of writing have been tools for management, administration and assessment. Those who can write can organize their collective affairs.

*Box 11.1. Self-assessment in other arenas:
Agricultural projects in the Republic of Mali*

The sort of self-evaluation methods outlined in this chapter are often quite applicable, with appropriate changes and adaptations, to enhancing popular participation in other areas of development. What is more, local people who have acquired valuable experience and "cut their teeth," so to speak, on participatory evaluation in the educational arena may prove both willing and able to undertake similar assessment responsibilities for assessing the results and methods of agricultural, health, natural resource management and industrial programs. The basic notions inculcated by the experience of participatory evaluation - identifying objectives and program "theories," finding simple ways to determine their degree of accomplishment, asking why the observed results have been obtained and not others, formulating suggestions on this basis for improved or more locally satisfactory performance in the future - are quite applicable across many domains. In addition, trainees of NFE programs with new literate or technical skills may provide just the local human resource that had been missing theretofore and so make it increasingly feasible for other development programs to turn their own necessary assessment efforts into locally-rooted exercises in participatory evaluation and planning.

The Associate Director of the national agricultural operation within which a good part of Mali's functional literacy program took place in the 1970s, and within which participatory evaluation procedures were first applied, strongly recommended as a result of that experience that the same approach be taken to evaluation each year of the agricultural campaign by development agencies, and that ample use be made of those newly literate villagers who had gone through the participatory evaluation of the NFE program. In the short run, his advice was not heeded; in the longer run, more participatory practices partly inspired by the evaluation have crept into agricultural and cooperative practice.

Reliability concerns

The potential weaknesses of this approach lie in the area of the reliability and comparability of results, and of their rather approximate or broad-gauged nature. The obvious complaint from a reliability standpoint is "Just how accurate can levels be if they are determined by semi-literate or newly literate adults themselves?" In fact, in most trials of the methodology learners who erred in their estimates were found to be

slightly more likely to *underestimate* than to *overestimate* their "true" level; and most hit the target pretty much square on the head.

But there are certainly possibilities of error throughout the procedure and this is one type of assessment where contextual factors play a necessary and important role. To minimize the possibilities of error, two types of cross-validation are typically used. The first, mentioned above, involves the local staff person (or outside evaluator) responsible for the assessment "verifying" the levels that learners attribute to themselves by proposing to them one or two other problems of the same level of difficulty drawn from a manual disseminated only to supervisory staff. This stratagem is, in addition, what makes it possible to disseminate the assessment brochure itself in advance of the evaluation date, so that learners have a chance to familiarize themselves with the procedure and to begin estimating their own levels.

The second method of cross-validation frequently used involves a "crosswalk" of staff from different localities a week or so after the initial assessment (long enough to avoid undue practice or "halo" effects, but short enough so that most learners will not have had a chance to learn a great deal more in the interim). In this formula, the staff person who performed the evaluation in locality A visits locality B, the staff person who performed the evaluation in locality B visits locality C, and so forth. In each case, the evaluator re-administers the test to a sample of learners and then compares this second series of results with the original ones.

The idea is not to revalidate every single learner's results nor to criticize individual staff for lapses in performance (though the process does generally prompt field personnel to greater care in their evaluation work), but rather to get an idea of the *margin of error* in the information overall. The magnitude of differences between levels determined by the first administration and those noted on the second administration furnishes a decent estimation of the degree of error in the assessment results, and one that can be used at all levels in interpreting the data.

Another area of potential weakness in this type of learner assessment is the rather approximate nature of results. Usually such self-assessment tools do not include more than four to seven levels. Beyond that, it may become difficult for learners to locate themselves on the scales. There are doubtless many more than seven dimensions and stages in the acquisition of highly functional levels of literacy, numeracy or life skills.³ The response to such concerns is twofold. First we return to the argument made earlier about the level of accuracy needed in assessment

data: very often even a reliable "order of magnitude" appreciation of learning outcomes that can be obtained locally by locality across the breadth of a program constitutes such an advance over currently available information that it would be frivolous to insist on a great deal more. Second, in the overall design of an evaluation program (discussed in Chapter Two) there are many reasons to include *both* a broad-based learner assessment of the type discussed here *and* a more refined analysis of achievement and learning problems which can be carried out in a sample of locations by direct intervention of qualified research or evaluation personnel.

In fact, this method of coupling self-assessment with external research has an additional payoff. Local staff and program participants who are actively involved in the self-assessment generally also become much more interested in the methodology and results of the sample studies, and an avenue is thus opened for maximizing the impact and utility of this outside research as well. The condition for realizing such synergy between outside research and learner self-assessment, one will remember from the discussion of participatory evaluation in Chapter Three, is respect of what we called the "concentric principle": both participatory evaluation and outside studies should be organized around a common set of questions and themes. If it is clear that the external evaluation is essentially designed to investigate in greater depth problems and issues posed by the participatory evaluation, then a basis is created for communication between the two and for increasing the active interest of staff and program beneficiaries in the external research.

Post-literacy assessment

Evaluating the "downstream" impacts of NFE and literacy programs is no easy task, as we will have occasion to say again in the next chapter, which is devoted to the question of impact assessment. Whereas one can be relatively certain that new levels of literacy or technical skill that learners manifest at various stages of an instructional program are directly attributable to the instructional intervention itself, this is not necessarily the case for broader personal and social impacts. A decrease in infant mortality in a village or neighborhood where there has been a health education program may be attributable in part to that program, but it may also stem from changes in the water supply, in-migration or out-

migration of families, vagaries of climate, or still other factors. Unraveling this tangled skein of cause and effect in order to say something (at least) about the broader effects of NFE programs is a major challenge for evaluation, and one to which we shall return.

But once again there is no point in throwing the baby out with the bath water, or refusing to do anything just because one cannot do everything. In fact, local self-assessment can add some very critical perspectives to the overall enterprise of determining NFE program impact. The key thing to remember is that "impact" in these situations is very largely a question of what happens to *people*, and the strength of a participatory evaluation is precisely that some of the key people are involved and can share, analyze and even debate their views of program impact.

The approach generally taken in participatory evaluation of program impact is therefore a modification of "tracer study" methodology. It typically involves at the outset the performance - with or by the learners and local authorities themselves - of a sort of "census" of graduates and former participants and an analysis of what they are presently doing, what applications they have found for lessons learned in the program and what obstacles they have encountered. One then seeks to interview - individually and/or collectively - a sample of these graduates in order to (a) retrace the "itinerary" (personal, educational and professional) that they have followed since leaving the program; (b) discuss with them the uses to which they have or have not been able to apply the lessons learned; and (c) ask them what changes in either the program or in the environment to which they returned would have ensured greater personal impact. One can even ask them to undergo the same kind of "learner self-assessment" (via the same simple brochure) that present participants in the program have undertaken, in order to get a general idea of the degree to which they have maintained, improved or lost their former level of competence.

Though the participation of former trainees or NFE "alumni" is obviously the essential ingredient in this sort of impact assessment, in fact the topics broached and the experience to be analyzed are typically of great interest to current students and other stakeholders of the program as well. Venues and methods that allow the entire group to discuss and analyze the "tracer" data are therefore very worthwhile. One approach is to organize a sort of "alumni day" at the training center devoted to attracting former students back to discuss their post-training experience.

The aggregation of this information can give local NFE program participants - and regional and central staff as well - a nice in-depth appreciation of the kinds of impact that the program is presently having and the reasons why it is not having more. As such, this dimension of the local "report card" or "balance sheet" is an absolutely critical element in the overall evaluation of an NFE program, because it touches on the reasons why people get involved in such programs and why local communities do or do not enthusiastically support them. The leading conclusion of a number of NFE program evaluations carried out in recent years has been "do post-literacy before literacy" (cf. Dumont 1986; Easton 1989). Self-assessment of these opportunities is therefore a necessary part of the determination and diagnosis of learning outcomes.

Conclusion

Self-directed learner assessment is thus not only a possibility, it is often an essential driving force in participatory evaluation of educational programs. Through it, program beneficiaries may gain habits and a level of confidence that will enable them to evaluate broader dimensions of the program. The practice of self-directed learner assessment also generally proves highly stimulating to the process of instruction itself. But much about the methodology remains to be refined and developed - and most of that refinement must take place in the field.

In this chapter, we have presented one possible methodology for enhancing learner participation in the determination of the instructional results or "learning outcomes" of NFE programs while at the same time safeguarding reasonable levels of reliability and validity in evaluation data. Technically, the approach is based on the broad body of theory about "criterion-referenced" testing, a branch of research and evaluation knowledge of particular relevance to nonformal education and training programs, where one is typically more concerned with assessing skills and competencies actually acquired and their practical applicability than with grades and the relative standing of different students. The added benefit of this self-assessment strategy is that it is potentially applicable to many of the other domains of local development in which nonformal educators are interested and in which their trainees or students may need to play important roles.

Notes

¹ For two useful references on the subject, see Lewis (1984) and Macgregor (1993).

² Shorck (1989) is one good text on criterion-referenced testing.

³ In addition, one can note, the method illustrated here assumes that knowledge acquisition can be laid out along single scales - i.e., that there is a natural sequencing in the learning process by virtue of which the learner will necessarily do A before doing B or C, and which can therefore be described as a series of distinct stages. Many knowledge domains of interest to NFE evaluators are not necessarily so strictly hierarchical: one doesn't always learn about the mechanisms of disease transmission before understanding something about human anatomy, and even in a basic skill like mathematics some learners may manage to resolve written or orally-posed problems before they master, say, the mechanisms of division.

Two responses can be made to this challenge. First, even those knowledge domains that are not inherently (or in any case univocally) linear and hierarchical are often presented that way in curricula and can be organized in that manner for testing purposes. Second, the self-assessment approach can also be based on more of a "matrix approach" to knowledge acquisition: i.e., one in which the evaluator defines a whole matrix or table of competencies and then helps the learner to discover which parts of this configuration have already been mastered.

References

- Dumont, Bernard 1990. "Post-Literacy: A Pre-Requisite for Literacy." In *Literacy Lessons*, an occasional publication of UNESCO/Paris. (Currently available as ERIC fiche number ED321058.)
- Easton, P.A. 1989. "Structuring Learning Environments: Lessons from the Experience of Post-Literacy Programs." *International Review of Education*. 35 (4): 423-444.
- Easton, P.A. 1983. "Functional Literacy in the West African Sahel: The Operation Arachide Project in Mali." In: J.Bock and G. Papiannidis eds., *Nonformal Education and National Development: A Critical Assessment of Policy, Research and Practice*. New York: Praeger.

Lewis, R. 1984. *How to Help Learners Assess their Progress: Writing Objectives, Self-Assessment Questions and Activities*. London: Council for Educational Technology.

Macgregor, J. 1993. *Student Self-Evaluation: Fostering Reflective Learning*. San Francisco, CA: Jossey-Bass.

Pollard, W.E. 1986. *Bayesian Statistics for Evaluation Research: An Introduction*. Beverly Hills, CA: Sage Publications.

Republic of Mali. 1978. *Evaluation de l'alphabétisation fonctionnelle dans la zone OACV* (Evaluation of Functional Literacy in the OACV Zone). Bamako: DNAFLA.

Shorck, S.A. 1989. *Criterion-Referenced Test Development: Technical and Legal Guidelines for Corporate Training*. Reading, Mass.: Addison-Wesley.

CHAPTER 12

IMPACT EVALUATION

We turn now to evaluation in the domain "downstream" from the instructional program itself: the impact that acquisition of this new knowledge, and participation in the NFE program *per se* has upon the well-being of the participants, their families and their communities.

What is meant by the term "impact evaluation"? What are its implications for and its applications to nonformal education? Definitionally speaking, the term "impact" refers to the effect that a phenomenon produces on something else, rather like a footprint in the sand or the mark left in a medium or material by some object. In educational evaluation parlance, the term "impact" is usually used to designate the effects of educational programs on neighboring domains of individual and social life - be they economic, social, political, cultural or psychological. Impact is thus something that takes place subsequent to, and as a consequence of, the educational intervention. In the CIPP/CIPOO model of educational evaluation presented in Chapter Four, a distinction of this sort is often made between a program's outputs (or its immediate learning results) and its outcomes (or the wider impact that it may have on its context and the personal lives of its participants).

Though impact thus represents what might be called a "second-order" effect of educational intervention, it nonetheless has a great deal of importance, particularly in situations - like those of most NFE programs - where the underlying goals of intervention are in fact not only to increase knowledge, but also to contribute to personal empowerment and to socio-economic, political and cultural development, however these are defined. One of the commonly repeated justifications for worldwide concern with "Education for All" is precisely that the spread of basic education is a necessary (if not sufficient) condition for the success of most other development efforts. So educational policy-makers and funders tend to be highly interested in the issue of program impact - and in assessing the cost at which it has been produced, a topic to be examined more closely in Chapter Thirteen. In this chapter, we will consider both ways for framing the investigation of program impact and a few of the techniques and methods for carrying it out.

Impact as a realm of program operation

The systems model that was introduced in Chapter Four also allows us to make another very important observation concerning program impact. Often in educational evaluations, program impact is treated as if it were simply a sort of metric for gauging the longer-term effects of the intervention rather than an integral part of the program itself. The systems model, however, suggests something significantly different and more in line, I think, with the actual experience of nonformal educators: namely, that socio-economic and cultural impacts are "produced" by activities flowing from (or sometimes occurring in tandem with) the instructional process and therefore constitute a domain of intervention as integral to program operations as administrative management or curricular design and delivery. Impact depends on the effective utilization of the knowledge acquired in NFE programs, and so on a series of application possibilities themselves contingent upon initiatives taken in (or conditions characteristic of) the surrounding environment. If the resources and possibilities for carrying on these "post-literacy" or post-training activities do not exist, then the programs probably will have correspondingly little impact.

Such a perspective also suggests that there is as important a needs assessment or "reconnaissance" activity to carry out in the outcome realm as in the teaching and learning one *per se*, and that ensuring positive impact may be just as dependent on careful programming as is success in instructional matters. There is, however, one critical difference: most programs in the post-literacy or post-training domain can only be carried out in collaboration with people competent in the other systems and practices that these activities involve: production cooperatives, new business or marketing ventures, public health programs, credit unions, newspaper publishing, natural resource management schemes and so forth.¹

Determination of impact is also a sensitive and sometimes contentious topic. For one thing, results in this domain are obviously less dependent on the quality of the educational intervention alone and more contingent on other resources and forces in the environment than is the case, say, for implementation targets like holding classes and ensuring good attendance, or even than instructional results like knowledge acquisition. It is a natural administrative and human reaction not to want to be held accountable for what one cannot control; and those responsi-

ble for educational programs may therefore approach the subject of impact assessment with some suspicion or trepidation. Like any topic that is both sensitive and highly important, impact determination needs thus to be dealt with diplomatically and skillfully by evaluators - but it does need to be dealt with.

In the rest of this chapter, we will consider some of the major steps and options in assessing the impact of NFE programs, from reconstitution of the program's strategy, through definition of potential impacts, to field inquiry and analysis of data. Given limited space, more emphasis will be put on conceptual than on technical issues, but the discussion should at least help practicing evaluators to locate areas of knowledge that they wish to develop further.

Reconstituting the program strategy

The first counsel for those undertaking impact assessment is so simple as to appear perfectly banal: a program cannot have an impact if it has not in fact been implemented! It is worth noting that the most common conclusion in evaluations of the series of educational innovations conducted in the United States in the later 1960s and early 1970s was that one could say nothing about the impact of the proposed methods because it turned out that they had in fact never been implemented as planned. (The evaluators did, though, have a thing or two to say about the administrative efficiency of the programs in question!)

This observation brings out a more general point - namely, that a program or project as actually implemented is seldom, if ever, identical to the official methodology of the project as initially planned. Strategies change form and are adapted to accommodate cultural, logistic and political considerations in the course of implementation. Ideally, this should be a conscious process of modification and rearticulation of the program strategy or program theory; but more often it is at least partly unconscious and undocumented, with the consequence that the methodology actually used in the field may differ significantly from the "official" one and will often vary appreciably from one place and from one period of time to another. Furthermore, bureaucratic or political obstacles may arise and stop implementation of a program in mid-course. Such happenings are frequent enough in North America that several books and numerous articles (e.g., Wholey 1987) have been

written on "evaluability assessment," meaning techniques for deciding whether a program has been sufficiently implemented to merit and benefit from an evaluation, or which of its components may be at this stage in its own evolution.

A prelude, therefore, to impact evaluation (and, really, to any evaluation) involves investigating these discrepancies between program theory and practice and trying to articulate the methodology (or methodologies) that are actually being implemented. As noted in Chapter Four, this exercise is itself a form of evaluation, since the modifications introduced in the strategy of the program in the course of implementation often reflect some experienced if tacit assessments of what will and will not work within existing political and economic constraints. Expressing these implicit lessons of experience can be an instructive undertaking.

Defining the program's effects

Once the evaluator has some handle on the program that is (or the various versions of the program that are) actually being implemented, then the inquiry into impact can get under way. The basic question involved in discovering the impact of a program is the "counter-factual" one: that is, how are the participants and the environment different from what they would have been had the program not taken place? Note that this is not quite the same thing as asking how the participants and the environment have changed since the inception of the program - because those changes may or may not be due (or may be only partly due) to the program itself. In fact, a program could have a positive impact even if the participants and their environment did not change at all! This might be true, for example, in a situation where economic or political circumstances were becoming more and more difficult or constraining. If, in the absence of the program, it seems likely that certain important conditions - such as literacy rates, maternal health, or farm productivity - would actually have gotten worse, then the fact that the program plausibly allowed them to remain stable could constitute a positive impact. In other words, though there was no apparent improvement, things actually changed for the better compared to what arguably would have happened if the program had not taken place.

Similarly, however, a program might have had no impact despite the fact that conditions improved during or shortly after its implementation,

if in fact that improvement would have come to pass even without the effects of the program - that is, if the favorable changes were basically due to other factors in areas outside the nonformal education effort. One could even imagine a program having a *negative* impact in a situation where related social indicators were *improving*, if conditions would have improved more in the absence of the program or had its resources been put to other uses!

So the underlying question remains, "How does the situation after the intervention differ from what it would have been had the program not been implemented?" This is actually a central and critical issue. Several things need to be said about it if we are to get to the root of impact assessment.

The fundamental uncertainty

First of all, no one, Almighty God Him/Herself excepted, really knows what would have happened if a given program had not been implemented, which is why this is referred to as the "counter-factual question." Nevertheless, the effort to distinguish the impact of the educational program from the effects of other concurrent social conditions is a useful exercise that does help a great deal in clarifying our thinking and orienting our evaluation of program strategy. In the more specific form often used in economic evaluations, the question is really how the situation differs from what it would have had the resources consumed by the program been put to other ends, and one proceeds then to consider other principal uses to which those resources might have been put and to compare their yield under the program with those potential yields.²

In short, one performs some sort of simulation in order to answer the "What if...?" question. Note that, whatever the theoretical pitfalls involved, most stakeholders of a program do this automatically and implicitly anyway: in other words, they ask and answer, to their own satisfaction at least, the question, "Was it worth it? Or would we have been better off to have done something different?"

Narrowing the focus

There are obviously hundreds of possible dimensions of program impact, just as there are innumerable facets of any social situation. Dr. Bhola details in his book a number of varieties of literacy program impact that

have proven relevant in past evaluations (Bhola 1990: 109-110). But if possible dimensions of program impact are very numerous - let alone innumerable - resources for evaluation are usually quite constrained: so how does one decide which sorts of impact to focus upon?

Selection criteria from program theory

There are various useful ways of narrowing the focus and *selecting particular aspects of potential program impact for evaluation*. The decision about which impacts to examine is generally made on a mix of theoretical, political and practical grounds. At the theoretical level, as noted in Chapter Four, the program "model" - that is, the set of objectives and methods that describe how the program is supposed to run - should suggest some important dimensions of impact. Evaluators in North America have taken in recent years to using the term "program theory" to designate the underlying assumptions about cause and effect - about what results are worth achieving, and about how they can be achieved and in what order - that underlie any social or educational program and that may be reformulated as it grows and gains experience. This model or theory, which is generally contained in documents on the program or, at least, in the heads of its designers and (hopefully) staff, should specify to some degree what impacts the program is designed to have and how educational changes are linked to other important social outcomes.

Of course, it's never quite that simple. Typically, goal statements for education programs are rather general and vague, and the exact strategy or interaction by which implementation of the program is to lead to given social, cultural or economic consequences is not very well laid out. For one thing, in the realm of social change, no one is exactly sure what works and methods must vary according to circumstances. On the other, as mentioned earlier, there is a natural administrative survival instinct at work here, because the less specific you are about goals and mechanisms for achieving impact, the less likely you are to be held responsible later on for not having achieved them. In addition, programs may be designed and implemented for a variety of personal and political reasons that have little to do (or at least little directly to do) with social development goals, and/or they may grow by fits and starts over a number of years, assimilating first one set of goals and then another one, as bureaucratic survival dictates. All this terrain may of course be mined with personal sensibili-

ties if not with actual political explosives and evaluators must know how to tread lightly without abandoning their quest.

Model specification exercise

For all these reasons, a program's goals and social impact strategy may appear to be a rather vague mishmash, though one that has had proven survival value. Such difficulties notwithstanding, the attempt to clarify impact strategy is well-founded, because it can reinforce a program's effectiveness in contributing to publicly-valued goals. Moreover, as the old adage says, if you don't know where you are going, you don't have very good chances of getting there. And you have even a worse chance of being able to assess your own progress toward that goal and to improve your practice correspondingly.

Therefore, evaluators may perform a real service by helping - and gently pushing - program designers and/or stakeholders to articulate just what impacts their intervention is designed to have and how these effects should be produced. Arguably, moreover, it is really inevitable and even desirable that initial goals be somewhat vague and incompletely specified, particularly if the project has been designed centrally, since one can't really fully define appropriate impacts and feasible mechanisms for accomplishing them before the program has begun to be implemented in the field and its supposed beneficiaries have been more fully consulted. Evaluation exercises can offer a natural and fruitful occasion to benefit from the experience acquired in order to get beyond the generalities about impacts that typically populate initial design documents.

The various methods that one can use to help program participants and staff elaborate or reconstitute the program model and thus specify priority impacts and their relation to each other are numerous, and they were discussed to some degree in Chapter Four. Suffice it here to say that (a) what is involved is basically a dialogue process where the evaluator(s) helps stakeholder groups to reflect on what the program should realistically accomplish and how; and that (b) the evaluator(s) should likewise anticipate both some differences of opinion among stakeholders (e.g., between central and field staff) regarding program methodology, and some friction and inertia, given the natural bureaucratic survival instincts cited above. In fact, organizational sociologists sometimes talk of the phenomenon of "goal displacement" - that is, the natural tendency of organizational staff to redefine their goals in order to correspond to whatever their programs have actually accomplished or are

able to do. (Aaron Wildavsky once labeled this syndrome in education programs "If we can't understand the questions, we must be taking the wrong test!" [1979: 106]) Despite these obstacles, the exercise of reflecting upon the nature of program impacts and the best method for achieving them can constitute a very valuable instance of organizational learning. Box 12.1 offers a concrete example of this exercise.

Considering unanticipated effects

Even if conducted intelligently and quite thoroughly, however, this reconstitution or elaboration of the program model would not suffice as the sole source of ideas about the potential impact of the program, because any social intervention can have important unexpected effects, as Michael Scriven (1973) is renowned for pointing out. An effort must therefore be made to brainstorm or inventory other possible effects, and then to select those that seem most plausible to the evaluators or are judged most important by stakeholder groups. Political and logistic factors will obviously come into play throughout this exercise. On the one hand, some subjects may be taboo in a given political environment or "too hot to handle" in other than indirect ways; on the other, some types of impact are more difficult to gauge than others, and the same practical and financial considerations that dictate limiting the number of dimensions of impact to be evaluated will incline the evaluators to limit the proportion of hard-to-research topics chosen.

Finally, note that causes and effects are typically linked in complex patterns. Quantitative researchers use the term "independent variable" to denote the supposed cause and "dependent variable" to denote the presumed effect. This terminology comes from the tradition of experimental studies, where the level of the independent variable is allowed (or made) to change and the impact that this manipulation produces on the dependent variable is then measured. In an experimental situation, the researcher tries to isolate the interaction of these factors, either by creating laboratory conditions or by careful structuring of experimental and control groups.

But in real life situations there are many other factors at play. Causal variables typically work through intervening factors and their effects are often masked or confounded by accessory variables. Figure 12.1 pictures just such a situation, where a training program produces an impact on agricultural productivity by influencing to some degree farmers' seeding methods and equipment purchase decisions. Yet these influences are not

*Box 12.1. Brainstorming program impact:
Assessing the effects of "service learning" in the University*

"Service learning" is a movement to involve students in community service not simply as volunteer workers but as an integral part of their program of studies and classroom learning. It amounts in effect to linking formal and nonformal education and harnessing the notion of experiential learning to that of community involvement with those in need. University students taking subjects from education through health to history or the physical sciences are involved first in using their newly-acquired knowledge to help develop and implement community projects or staff local schools and training centers, and then in analyzing and learning from the results of these initiatives.

In the mid-1990s, funders of such programs in the United States asked for evaluations of the "impact" of the activities. The first question for evaluators was "what impacts should one expect or look for?" Service learning was a very diverse movement, taking varying forms in different locations, with many values in common but no single "program model." The evaluation team spent some time "brainstorming" possible impacts with staff and participants in service learning programs and soon came up with a simple three-by-two grid as a framework for the exercise:

	Anticipated Effects	Unanticipated Effects
Personal Impacts		
Institutional Impacts		
Community Impacts		

"Institutional impacts" referred to effects on policy and practices within the University itself, "personal impacts" to individual effects on participating students, faculty and beneficiaries, "community effects" to changes in local social conditions. This grid provided a tool for discussing both anticipated and unanticipated effects with program staff. In order better to begin detecting unanticipated effects, the evaluators conducted a first set of open-ended interviews with a few participants and key stakeholders to get a sense of the meanings and impacts they attributed to the experience.

direct and are modified by a host of accessory variables in the environment, like annual rainfall and the availability of new farm equipment. Of course, Figure 12.1 is itself a highly simplified and schematized version of reality, but schematics like this may be useful in framing impact evaluations.

Gauging actual impacts: Rationalistic and naturalistic approaches

Once dimensions of impact are identified and operationally defined, how does an evaluation team go about assessing the actual incidence of these effects - or of others as yet unsuspected? In fact, there are both quantitative and qualitative approaches to this issue (or rationalistic and naturalistic evaluation strategies for addressing it), and in most situations the greatest payoff comes from combining them in some fashion and ensuring a good deal of participatory analysis of results. Let's first look briefly at the two types of approaches

Quantitative or rationalistic strategies

Rationalistic approaches to the determination of program impact basically involve collecting quantifiable data in order to analyze the degree of correlation between measures of program implementation on the one hand and, on the other, indicators of change in the various characteristics of participants or their environment (like their income, the health of their families, the financial stability of agricultural credit unions to which they belong, or the demonstrable militancy of their political organizations). The indicators chosen are those deemed to reflect important dimensions of the program's impact. In the tradition of scientific research, four things are required to demonstrate a causal relationship between two phenomena (which we will call "A" and "B" for simplicity's sake):

1. There must be good theoretical and/or practical reasons for suspecting that A is the cause of B.
2. The occurrence of A must precede the occurrence of B. (If the lightning happened after the fire started, it is difficult to maintain that the lightning caused the fire!)

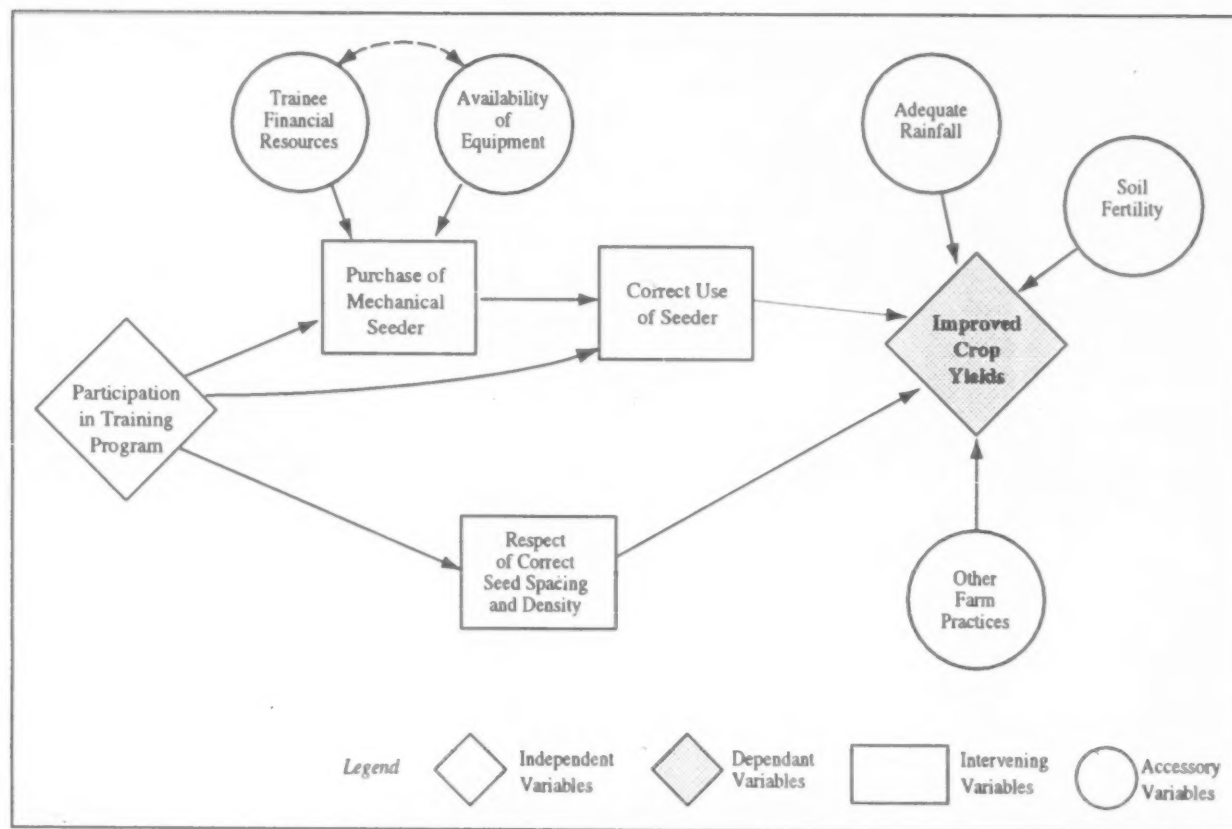


Fig. 12.1 Casual scheme for relations between training and improved agricultural yields.

3. The intensity of the two phenomena should be "correlated" or "associated" - that is, the more A is present in a given environment, the more one should find B, other things being equal...and inversely.
4. Finally, one must have some way of either excluding the influence of factors other than A that could cause B to happen (which is done in experimental research by creating experimental and control groups) or of parceling out the role of the other factors in causing phenomenon B (which is the goal of "quasi-experimental research"). In quasi-experimental research, one uses statistical analysis of variations in the presence or intensity of A, B and other specified factors to try to identify the role or weight of each in producing B.

One never satisfies these four conditions perfectly. There are weaknesses in the most careful experimental or quasi-experimental research. Moreover, as we shall see again in a moment, our ambitions in evaluation are often more modest and immediately practical than is the case in social and educational research *per se*. Nevertheless, evaluators will often use certain strategies and elements of quantitative and causal research in developing their investigation of program impacts - for example, by employing cross-tabulation tables or correlation coefficients to examine (at either the individual or the collective level) the association between degree of success in implementing the program and the prevalence or intensity of different indicators of impact. It is possible without great difficulty to construct these analyses in such a way that one at least "controls" for the influence of one or two other possible intervening factors. For example, if gender may make a difference in the impact of NFE lessons on individual health practices, then the analysis of correlation between exposure to NFE and modifications in health-related behavior can be performed separately, once with data exclusively on women, and the second time with data exclusively on men. Further controls for accessibility of pharmacies or level of personal income can also be introduced if the data are available and the sample is sufficiently large to support this kind of disaggregation.

Qualitative or naturalistic strategies

Chapters Eight and Nine presented some of the main characteristics and approaches of qualitative evaluation: methods like case studies, participant observation, and in-depth interviewing. What are naturalistic

strategies of *causal investigation and analysis*? Essentially, in the qualitative mode of evaluation (or research) one makes use of the "human instrument" - whether the evaluator her- or himself or key informants and interviewees - to gather and synthesize a much deeper and more "dimensional" set of information about situations of interest than one could easily get through quantified indicators and (e.g.) closed questionnaires.

Qualitative analysis of causation involves investigating what might be called the "micro-mediation" between cause and effect - that is, the detailed sequencing and meshing of phenomena that describe how events lead to and are connected with each other. Patton (1990, p. 147) uses a cartoon to illustrate this concern for intervening detail. It depicts a journalist (the qualitative evaluator) interviewing a star football player (the quantitative one). The journalist says, "Last year in the opening round of national cup competition you scored five goals and this year you only scored one. What was the difference?" And the player responds, "Four"!

This is obviously a bit of a caricature, since true sports analysts would use many other statistics in analyzing the problem: average goals scored against opposing teams, numbers of assists made by the player in question, and so forth. But the point is well taken. Fully understanding the drop in performance would require talking in greater depth with a number of people, better grasping what the player is going through personally and professionally, experiencing the ambiance and dynamics of the player's team and the conditions on the field, and appreciating ways in which these may have changed over the intervening year. Only in this way could one reconstitute something of the web of influences that shapes performance on the field. Of course, no one can absorb all the experience and data that this might imply: qualitative inquiry, like quantitative, is confronted with a question of selection or sampling. But the instinct of the qualitative evaluator is less to study correlations between a few schematic factors on the input and output sides of experience than to portray the process that intervenes between them.

There is, in fact, a special relevance of this approach to causal analysis: One problem with even the best experimental or quasi-experimental studies is that the investigator has to use quantified indicators of the causal factors in question and - because of the difficulty of quantifying the related concepts - these are often fairly removed from the

subjects that one really wants to investigate. At the same time, the indicators of cause and effect are sufficiently removed from each other so that the results, even if they suggest a causal impact of A upon B, give little idea of just how it might have come about. Suppose our indicator of success in an NFE program is the numerical results of a rather inadequate test and our indicator of health practices is simply the number of times that former participants bought malarial prophylaxis. (Practical constraints often lead to the adoption of such indicators in quantitative studies.) Under these circumstances, even if the two indicators turn out to be highly (and significantly) correlated, it is questionable to what degree we can conclude that participation in the NFE program really changed health practices; and in any case we still understand very little about what elements of the NFE experience had this effect, and so very little about how to consolidate this aspect of its impact.

This example illustrates how useful qualitative or naturalistic studies can be in impact analysis. First, a qualitative investigator can develop a number of ways of perceiving the level of genuine participation of people in NFE classes, can potentially get a much fuller impression of the health practices of the individuals concerned, and can even gain a new appreciation of what "health" and "sickness" mean for them and to what aspects of the course they attribute whatever impact they feel that the experience had upon them. All this can result in greater validity of conclusions - that is, greater confidence that one is actually dealing with the concepts of "NFE participation" and "effective health practice" that are of real interest and with their relationship, rather than just with some indirectly related numbers of doubtful validity.

Second, by careful naturalistic observation or interviewing, one has much better chances of reconstituting a realistic network or sequence of factors that link NFE participation to health practice or that modify the relationship between the two - including gender effects, the influence of poverty, the messages simultaneously being transmitted by the available media, and so forth. In the real world, causal relationships are not schematic correlations of a single factor A with another called B, but a complex interaction of many events and forces with a series of intermediate steps that lead from the cause to the supposed effect. Naturalistic evaluations are particularly well-suited to picking up this web of relationships and understanding something of its complexity. Quantitative methods may then sometimes be used to "test" the relationships discovered in this way. Ironically, the quantitative exercise is useful as much

for the fact that it forces the evaluator to specify and map the factors involved as for any additional certainty - or uncertainty - that its statistical results add to conclusions.

In any case, qualitative casual analysis is a natural and necessary part of the investigation of program impact. Lawyers, parents and politicians use it extensively. Only social researchers seem not to have caught on until quite recently.

Mixed methods and triangulation

It should be clear from the foregoing discussion that evaluators generally benefit most from a careful combining or "triangulation" of qualitative and quantitative approaches, as the strength of each one tends to compensate for the weaknesses of the other. Quantitative approaches are easier to apply to large populations or samples and (providing one has some initial framework) easier to analyze, whereas qualitative approaches are more limited in geographical scope and take a bit more time to analyze. Quantitative methods entail more rigorous and precise preliminary definition of the factors and forces involved, which is both a virtue and a limitation. At the same time, qualitative methods give a deeper picture of how impacts are produced and what they mean to participants, without which statistical data have little meaning. Overall, both methodologies deserve to be carefully applied, but can be used at varying levels of refinement. How to mix and combine the two is an issue with which evaluators do well to accumulate as much experience as possible. Apprenticeship, practice and critical evaluation of one's own work - as well as (in certain circumstances) formal training - seem to be the keys to growing competence in this domain.

"Tracer studies" constitute a particularly good example of an impact evaluation strategy that combines qualitative and quantitative approaches (cf. Zideman 1995, Carnoy 1988). This technique is conceptually quite simple and highly informative. It can be logistically difficult to realize on a large scale, but small scale applications are both possible and very informative. As the name implies, these studies basically involve following up on former program beneficiaries to see what they have done and become since completion of their training, and what they now feel about the effect that it had upon their lives.

Tracer studies can have an important quantitative dimension, which lies in the measurement and compilation of changes that have occurred

*Box 12.2: Tracer studies - no problem:
The census of new literates in Niger*

"Tracer studies" are types of evaluation research designed to elucidate what has become of participants in a program and what they have done with the skills or knowledge acquired. The methodology can be quite complex and statistical, but it need not necessarily be so. In 1980 personnel from the Literacy Agents Training Center (CFCA) in Niamey, Niger undertook a tracer study of former adult literacy students as one means of gauging the impact of that country's literacy campaigns and of better understanding the applicability of new literate competence in a rural environment.

The study lasted two months and entailed visiting a sample of villages in the western region of the country to perform, with local officials and literacy committees, a "census" of the whereabouts and current occupations of adults who had completed the literacy program over the previous ten years, then interviewing a subsample of these - including a few no longer in the village who could still be located - and testing their current level of literate competence with a simple self-evaluation form. An inquiry guide, questionnaire and self-evaluation form were prepared for the study and pre-tested in five villages before being revised and finalized. The study involved collecting both quantitative (e.g., rates of out-migration and of literacy retention) and qualitative (e.g., in-depth interviews regarding the uses of literacy) data.

in the income levels, social status, or other competencies and habits of a relatively large group of program graduates, compared sometimes to changes in the lives of a similar group of individuals who did not participate in the program. But they also have a very important qualitative dimension, which generally consists of reconstituting with a sample of graduates the "itinerary" that they have followed since program completion, hearing their stories and exploring with them the factors that have enabled them to use - or have prevented them from using - the skills and knowledge acquired.³ Box 12.2 provides an example.

Organizing impact evaluation

How can impact evaluation best be organized and integrated with the other tasks that evaluators face? Whatever the importance of this aspect

of NFE assessment, if it can't easily be related to the rest of the responsibilities NFE evaluators and sponsoring agencies typically must shoulder, it is likely to remain at best a laudable intention.

Several considerations are worth noting in this regard: the natural linkage between impact assessment and post-literacy evaluation; its close relation with management information system (MIS) concerns; the importance (to decision-makers in particular) of "prospective" impact evaluation and of economic varieties of impact assessment; and the great benefit that can be drawn from intersectoral collaboration in this domain.

The relationship between impact and post-literacy

To begin with, in NFE and literacy programs, the focus of impact evaluation coincides to a considerable degree with the "post-literacy" or post-training domain, since the impact of an instructional program mostly occurs after the program has taken place or among those who have already taken part, and it concerns what they are able to *do* with their newly-acquired knowledge, skills and attitudes (and how they feel about these results of their training).

Let's remember, however, that impact evaluation deals with *what-ever* changes take place in the comportment of former NFE participants and the conditions of their environment as a consequence - or at least a partial consequence - of the instructional program. It is *not* concerned purely and simply with the conduct and results of the particular post-literacy activities organized by NFE agencies, as important as these are. It is equally if not more critical to assess changes that come about independently of further programmatic interventions by the promoters of NFE. It is also important to note that we are talking not simply about the continuing education, reading and literacy retention of former participants, as important as these are, but equally about the applications of their new competence to other areas of life and the consequent gains in personal satisfaction and/or productivity.¹

Second, some impacts may occur in the environment well before the program has run its course, and some may even be produced by the very establishment of the program: one might call these "concurrent impacts." The program itself constitutes a specific kind of organism inserted into the existing urban or rural environment or grafted onto its institutions, and it can be worthwhile to assess how other systems and people in the environment have reacted to its operations and modified their own

behaviors as a consequence, for these themselves are instances of impact.

The importance of good MIS data

Though the sort of ongoing program monitoring that management information systems are designed to facilitate may seem at quite some conceptual and temporal remove from the assessment of the long-term impacts of the intervention, the latter is in fact highly dependent on the former. Decent investigation of impact generally requires *longitudinal* data - that is, information on participants, instructional results, and environing conditions at various points in time, not just at the precise and particular period when the evaluation exercise is undertaken. Without such benchmarks, change cannot be easily assessed.

This kind of data can only be furnished in a durable way by some portion of the management information system. It is thus quite important to reflect on key indicators of impact when devising an MIS, so that they can be introduced into the strategy of data collection at that point, or at least at some subsequent point well before analysis of impacts is required. This practice has, moreover, the added virtue of giving the users of MIS data - decision-makers, in particular - a means and a reason for staying abreast of what is happening in the environment of the program and to its "graduates" and thus an impetus to keep the long-term objectives of the intervention firmly in mind.

Practical applications: "Prospective" and cost-benefit approaches

At the same time, decision-makers are more likely to take impact assessment seriously if they see its relation to other activities of practical interest. Two areas where this focus on long-term outcomes overlaps with operational concerns of program management are therefore worth highlighting: feasibility studies and calibration of program benefits.

Feasibility studies

Impact evaluation can be *prospective* as well as *retrospective* and, in that form, is very closely related to planning. Estimating the wider impacts that an educational intervention may have is often an important part of program or project design, and (prospective) "social impact analysis" is frequently undertaken in domains as diverse as environ-

mental planning and industrial development. The habit of investigating - quantitatively and qualitatively - the broader impacts of NFE programs is probably the best preparation for doing this sort of prospective analysis. Such efforts are often referred to as "feasibility studies" (cf. Gray 1982): investigations carried on before program initiation in the field to make some estimate of likely success and judge whether the probable results will be worth the resource investment.

Though the term "feasibility studies" is usually associated with funding decisions by major institutions, this technique of prospective impact evaluation is in fact quite relevant to local NFE programs and individual participants. The Tostan nonformal education program in Senegal includes as the last step in its basic rural training sequence - targeted predominantly, but not exclusively, at women - a module on feasibility analysis designed to enable participants to apply their new skills to identification of small business and community development opportunities and *ex ante* assessment of their potential viability.

Considering economic impacts

Impact evaluation is also one important part of "cost-benefit analysis." The other part evidently is the determination of unit costs. The ultimate concern of economic evaluators is to compare the demonstrable impacts of a program to the resources that had to be sacrificed in order to produce those benefits. The validity of their analysis is therefore in large measure dependent on the degree to which they can, in fact, identify and assess the principal outcomes of a program, or base their work on impact evaluations already established by others - and the survival of programs supported by outside funding is, in turn, often dependent on how convincingly they can demonstrate the benefits that they have produced. In the next chapter, we will look more closely at the methods that economic evaluators use in their examination of NFE programs. Note in passing, however, that cost-benefit analysis may also be prospective and, in this form, is essentially similar to the evaluation that lending institutions make of loan requests - or the analysis that Tostan participants learn to undertake before launching their own small businesses and community development projects. It constitutes a good exercise for anyone interested in project or program planning, because the effort to quantify possible costs and impacts brings the likely success of the initiative much more clearly into view.

Intersectoral collaboration in impact assessment

Precisely because the larger impacts of literacy and nonformal education programs are so contingent on other factors and other programs in the domains of economic development, health, political organization and so forth, impact evaluation presents an ideal occasion for *intersectoral collaboration*. NFE, literacy and training evaluators will be well advised to attempt coordinating their work with that of evaluation personnel from related agencies in the Ministries of Health, of Rural Development, of Community Service, and of Agriculture as well as those of relevant NGOs. There is a great deal to be gained from establishing channels of communication across these administrative divides, exchanging ideas and experiences about how best to evaluate social impacts, working on blending evaluation and research studies whenever feasible, and generally contributing to consolidation of the larger "evaluation profession" in the country in question.

At the same time, the understanding of the relationship of NFE programs to other development activities and public policies in their environment - which impact evaluation brings into much clearer focus - can be very helpful in suggesting directions for post-literacy or post-training effort. In exploring these intersectoral connections, one is generally impelled to begin mapping the whole network of linkages and external factors that determine possibilities for effective utilization of new literate and technical competence in the environment, and so to discern the sorts of policy changes that might make this environment more favorable to adult learning and that might be targeted by post-literacy initiatives. Moreover, even if a cross-sectoral evaluation suggests that the external impact of NFE programs has to date been minimal, it also offers an excellent occasion to analyze *why* this is the case and why such an important human resource has not been put to better use. That may lead in turn to identifying policies and measures that would raise the value of nonformal education and increase the benefit that participants and their families, workplaces and communities draw from it. Those sorts of policies and measures are typically cross-sectoral concerns, for they inevitably associate literacy, training and nonformal education with the different domains in which this new knowledge is used. Diagnosing the blockages that impede fuller use of new skills by different population groups is an exercise worth sharing.

Conclusion: Participatory evaluation of impact

Though impact evaluation is related in many ways to social science research, it is also a domain where the differences between evaluation and research may appear quite clearly, particularly as concerns the level of causal analysis undertaken. In research studies of social change, the researcher often goes to some length - in either a rationalistic or a naturalistic mode - to retrace a network of causal connections and to estimate the relative influence of a number of different factors. These are exercises that require a great deal of patience and persistence and some technical finesse. Evaluators are typically less concerned with establishing the exact structure of causality than with judging to what degree the project or program can be said to have had certain stated effects and what unanticipated consequences may have resulted. (In fact, from a political perspective it is often said that correlation *is* causation in the evaluation lexicon - because program managers will be quite eager to claim credit for any positive changes that may have occurred concurrent with their intervention, whether demonstrably linked to it in a causal sense or not, and, as long as constituents and intended beneficiaries are relatively content, politicians may be quite willing to let them do so.)

One should never consider impact evaluation as necessarily or exclusively a sophisticated realm of inquiry only to be undertaken by professional researchers: nothing could be farther from the truth. It is, in essence, another one of those topics that are too important to be left to the specialists. While various types of rationalistic and naturalistic methodology can be very helpful in deepening and broadening this dimension of NFE evaluation, the core questions are simple ones; and a very useful impact assessment can be conducted simply by asking former participants what they have managed to do with their new competencies, what obstacles they have encountered in this process and what worth these results have for them.

"Keep it simple", "make it participatory" and "just do it!" may therefore be appropriate notes to sound in closing this chapter. Impact evaluation needs to be a highly participatory type of evaluation, because it is impossible to get a clear idea of the effects of NFE programs on their participants without eliciting their own analysis of the subject. The qualitative side of tracer studies discussed above illustrates one example of this principle, but its applications are in fact a good deal broader.

Typically participatory evaluation at the local level focuses considerable attention on impacts, because the "utility" of the program (from cultural and psychological as well as economic points of view) is something of live and immediate interest to local stakeholders, who happily are not inhibited from analyzing this question by the fact that they do not have a command of multivariate statistics or a string of university degrees! All of us *think* in cause-and-effect terms, even if we do not master formal methodologies for investigating such relationships. And any move towards democratic accountability virtually requires that stakeholders be enabled and allowed (i.e. "empowered") to judge what results may be validly attributed to an institution's ministrations and whether those outcomes were worth the effort. This chapter has addressed the first part of that challenge. Chapter Thirteen will address the second.

Notes

¹ In the literature on literacy and in the practice of the agencies and organizations sponsoring programs, "post-literacy" is sometimes taken to refer strictly to the initiatives made by literacy personnel themselves to provide post-training application opportunities for program participants: essentially, development and supply of reading materials, creation of local libraries and newspapers, and occasionally further cycles of training - in short, the kinds of things that educational sponsors can provide, given adequate resources. From the perspective of impact assessment, however, it is evident that this is only one side of the coin, and that in the absence of initiatives on the other side - namely, in the socio-economic and cultural realms where literacy might be used as a tool - those continuing education programs risk not being very sustainable.

During the evaluation of nonformal education in Mali in the late 1970s, researchers developed a simple two-by-two grid to portray the multiple dimensions of post-literacy and their interdependence. It is pictured in Figure 12.1 hereafter.

² For orthodox neoclassical economists - that is, believers in pure free market capitalism - the productivity of resources in all possible alternate uses is indicated by their competitive market price. Logically, if all potential users of a resource have the opportunity and the wherewithal to bid upon it, then the one who gets it will be the one best able to employ it, and the price she or he pays should approximate its true marginal productivity. This is the basis of much conventional cost-benefit analysis. The assumptions come shining through.

Figure 12.1 Multiple dimensions of post-literacy

		Post-literacy initiatives	
		Continuing education	Socio-economic functions
Parties responsible	Appropriate agency sponsors	<i>Advanced training sessions; access to other forms of education</i>	<i>Administrative and socio-economic decentralization; creation of new local functions and responsibilities</i>
	Local communities	<i>Self-instruction and lifelong learning</i>	<i>Local community development initiatives</i>

³ Sampling and representativity concerns are, however, frequent in tracer studies. The problem is how to gather data on a representative group of former students. Some are very difficult to contact, and these people may be systematically different from those on whom information is most easily available.

References

- Bhola, H. 1990. *Evaluation "Literacy for Development" Projects, Programs, and Campaigns: Evaluation Planning, Design and Implementation, and Utilization of Evaluation Results*. Germany: UNESCO Institute for Education.
- Carnoy, M. 1979. "Education and Employment: A Methodology for Local Policy Research." *UNESCO S.71*.

Gray, P.J. 1982. "A Guidebook for Conducting Field Trials of New Methods." Paper and Report Series No. 77.

Mohr, L.B. 1995. *Impact Analysis for Program Evaluation*. 2nd edition. Thousand Oaks, CA: Sage Publications.

Schalock, R.L. 1995. *Outcome-Based Evaluation*. New York: Plenum Press.

Patton, M.Q. 1990. *Qualitative Evaluation and Research Methods*. Newbury Park, CA: Sage Publications.

Scriven, M. 1973. "Goal-Free Evaluation" In: E.R. House ed. *School Evaluation: The Politics and Process*. (319-328). Berkeley, CA: McCutchan.

Wholey, J.S. 1987. "Evaluability Assessment: Developing Program Theory." *New Directions for Program Evaluation* 33: 77-92.

Wildavsky, A. 1979. *Speaking Truth to Power: The Art and Craft of Policy Analysis*. Boston: Little Brown Company.

Zideman, A. 1995. "Many Paths to Skilled Employment: A Reverse Tracer Study of Seven Occupations in Colombia." *Education Economics* 3: 61-79.

CHAPTER 13

ECONOMIC EVALUATION AND COST-BENEFIT ANALYSIS

Impact studies focus naturally on the *results* of NFE programs. But policy issues cannot be resolved nor can planning decisions be made without reference as well to the *costs* of these programs. However great their anticipated benefits, if the costs of such efforts are prohibitively high or cannot be offset by realizable benefits or cost-recovery mechanisms, the effort

simply will not be feasible. There might be great demonstrable benefit, for example, in implementing fully interactive computer instruction in every adult literacy center or vocational-technical training facility throughout rural areas of a given country, but the expenses involved may well render this idea impracticable, at least in the short run. All in all, available funds might be better spent on other things.

If impact studies and analyses of the effectiveness, quality and relevance of NFE programs are to be of use to policy-makers, they must at some point be combined with data on costs. To that end, costs must be carefully determined and measured: operating costs and investment costs to each of the parties concerned by a given program. Once data on expenditures are assembled and aligned with effectiveness measures, ways must be found to compare the two types of information and to draw conclusions. This chapter is devoted to examining such "cost-benefit"¹ measures and the reasoning that underlies them.

The language and the reasoning may be unfamiliar and even slightly repugnant to some readers. By virtue of their value systems, educators often have a certain degree of allergy to economic concerns and therefore correspondingly little familiarity with them. In addition, economists have generally done a good job of surrounding the area with enough mathematical terminology to repel those who feel they lack numerical genes. This is unfortunate. On one hand, the basic ideas of economics do not require extensive mathematical treatment, and the numerical conventions used are really quite simple. On the other hand, economics is an area of too much importance - for the

future of education as well as for other social domains - to be left to the economists. For this reason, and on the strength of the adult education principle that the only useful knowledge is shared knowledge, readers are encouraged to gather all their resolve and to forge ahead with the presentations on the following pages. We shall look successively at a few of the underlying notions of educational economics; some methods for defining and assessing the costs and benefits of programs; the different levels of economic analysis that may be applied to the evaluation effort; methods for comparing costs and benefits; and the implications of these varied techniques and approaches for nonformal education.

The economics of education: Underlying notions

What does the field of economics have to offer to the concerns of NFE evaluation? Economics has been variously defined, but a typical description from an introductory text suggests that it is

the study of how people and society choose to employ scarce resources that could have alternate uses in order to produce various commodities and to distribute them for consumption, now and in the future, among various persons and groups in society (Samuelson & Nordhaus 1985: 4).

More succinctly, Lionel Robbins calls it the study of "human behavior as a relationship between ends and scarce means that have alternate uses" (Robbins 1963: 23).

Among the "scarce resources" of interest to economists are human ones - the skills, motivation, insight, energies and labor of people. These are an integral part of any economic activity; and though there have been long debates about how to understand and represent the role of labor in production, "human resource economics" has in the last thirty or forty years become a very important subfield of the discipline. Human resource economists - or economists of education, as those specifically concerned with phenomena of schooling and training are called - focus on the production and exchange of the complex set of services involved in organized facilitation of learning.

They therefore analyze topics like the supply and demand for different types of educational services, the costs and benefits associated with producing them, the way in which the necessary social transactions are structured, and the impact of these activities on the welfare of different individuals and social groups.

There is much here that is of potential interest to organizers and evaluators of nonformal education, though it is couched in a language with which they may not be very familiar. Before turning to some of the mechanics of economic evaluation of education and training programs, therefore, let's take a moment to examine five core notions of economics and the terminology in which they are commonly expressed: resources, scarcity, costs and benefits, supply and demand, and efficiency.

The notion of resources

Economics, as is evident from the definitions above, has a lot to do with the generation, allocation and use of resources. But what are "resources"? This broad term refers in economic parlance to anything that enters into human production or consumption, anything that can be used for the perpetuation and adornment of human life and that is consumed or transformed in the process. Ecologically-minded economists would want to broaden the notion still further to include whatever serves for the perpetuation and adornment of the biosphere in general, though to the degree that dependence of the human species on its natural environment is acknowledged the first definition sooner or later implies the second.

This notion therefore includes "natural resources" (or those factors of production and consumption like land, sunlight, original forests and mineral deposits that existed independently of human intervention), the "human resources" already mentioned, and "produced resources," or results of human productive activity like machines, cultivated crops, and cultural artifacts. The cake can be cut in a number of ways, but the common denominator of all these items is that they can serve as inputs to other productive processes or as objects of final human consumption.

Education itself is a type of human resource, since it enters both into future production activities by potentially increasing the skills of

labor and into future consumption as a source of personal enjoyment and cultural enrichment for those who have benefited from it. In fact, economists regularly treat education as a fundamental aspect of what they call "human capital," or all those attributes that people inherit or acquire which later pay dividends to them and society at large by enhancing their productivity and/or income, a line of thinking first formulated in full dress by Becker (1964). Less regularly they admit that education also has the nature of what is called in economics a "consumer durable" - that is, something acquired for its personal, rather than market utility, which provides long-lasting but non-salable benefits to its owner.

The notion of scarcity

The definitions of economics quoted above also reveal the emphasis that economists put on notions of scarcity. Resources, they insist, are not limitless and - a few pashas, industrial barons and Microsoft officials excepted - no one and no society can simultaneously satisfy all its needs. The problem that immediately arises, therefore, is one of *allocation and choice*. To what should we devote our resources, given that they do not suffice to meet all our needs at once? How do we identify those resource uses that will be the most productive for us, through either increasing our future holdings or more effectively satisfying our present wants and needs? How do we choose among the *alternate uses* that such resources have? Economics purports to help us answer such questions and has even been grandiosely termed (pun certainly intended) "the science of choice."

Devoting resources to one use automatically means that they cannot be allocated to another and that we will have to forego whatever that alternate use of them might have furnished to us. "You can't have your cake and eat it too," to quote the old adage. As we will see more fully in a moment, therefore, the fundamental type of cost in economics is what economists call an "opportunity cost" - that is, the value of the foregone uses of the items in question. From this perspective, for example, volunteer labor is never free, at least not to "society as a whole" (another concept to be examined in a moment), because the volunteers in question could have spent the same time in some alternate use of their talents, perhaps remunerated, and the cost to

society of the labor they contribute in voluntary endeavor is equal to its worth in its best alternative use. This emphasis on scarcity and hard choices has given economics the nickname of "the dismal science", as well as a reputation for rigor. Neither is entirely deserved, but economists are fundamentally concerned with tradeoffs and comparative advantage.

The notion of costs and benefits

What are the "benefits" and "costs" of an educational program? The meanings given these terms in economics are a little different from common sense perceptions of them. A benefit is basically defined as anything that confers some perceptible utility on the person who acquires it - by partially or wholly fulfilling a felt need, by generating pleasure or satisfaction, etc. A cost, on the other hand, is simply a foregone benefit. We incur a cost whenever we have to give up a resource that, used in some other way, might have procured us desired benefits. The traffic fine "cost" me a certain amount because, in order to pay it, I had to expend money that could have been used to purchase a new suit and spend time that I could have devoted to my children.

The underlying perspective on resource expenditure used here comes from the notion of "opportunity costs" reviewed above. Opportunity costs need not involve any obvious cash transaction, since they are basically an estimation of missed productivity. In fact, in economics all prices and wages are theoretically based on this same notion: in a competitive situation, to hire someone or buy some good, one is obliged to offer at least as much as that person or item could have commanded in its best alternative employment.

In identifying benefits and costs, we are trying to understand how best to use or "allocate" the resources at our disposal, in order both to meet our own needs and to generate additional resources for the future. The assumption of scarcity therefore underlies this effort as well.

Notions of supply and demand

Concepts of costs and benefits, consumer choice and allocational decisions only make sense in a framework of exchange and social

interaction. Economics is, after all, one of the "social sciences" and the basic paradigm underlying economic thought is that of free market exchange: the organized and yet largely unfettered commerce of goods and services. The two driving forces in the free market paradigm are supply and demand. The owner of useful goods and services exchanges them with the consumer for an amount of currency or other resources set, in principle, by an interaction between the relative scarcity of the items in question and the relative urgency of the needs which the consumer wishes to satisfy by their acquisition. In economic parlance, supply interacts with demand: sellers and buyers seek an equilibrium position where the price at which the former is willing to cede the desired goods or services matches what the buyer is willing or able to spend in order to obtain them. Then, and only then, can exchange proceed.

An economy is made up of a linked series of markets that provide for the exchange of products and services among primary, intermediary and final producers on the one hand and between them and final consumers on the other. This chain of exchange activities stretches from the extraction or mobilization of the original natural and human resources required for production through the various stages in the processing of desired goods and services to their final marketing and consumption. From the economic point of view, organized learning of whatever variety plays an important role in two of these arenas of exchange - the market for educational services and the market for trained labor - and it is linked in some sense to all the others.

In the market for educational services, suppliers of schooling and training meet potential consumers, namely future students and their families. Where the resources that the latter are willing to expend for instruction match the prices that the former are willing to accept for their services - plus the opportunity costs that students necessarily incur in devoting time to their education - the transaction is consummated and the person engages or enrolls in the learning sequence in question.² The prices asked by suppliers of educational services are obviously related to the costs of producing ones of acceptable quality, and therefore determined in turn by teaching salaries, infrastructure costs and so forth. As for consumer willingness-to-pay in this market, it is often referred to as a "derived demand": that is, the level of interest in a given educational experience is directly related to (if not

uniquely determined by) the demand in labor markets for the kinds of skills and knowledge that the education in question reliably provides or the amount by which this training sequence will plausibly increase the productivity and/or earning power of the recipient.

The two markets are thus closely linked in economists' eyes. The consumers in the market for educational services - i.e., the students - become the suppliers in the labor market. They are now offering to potential employers (including themselves, should they choose to go into business on their own) a set of skills and competencies that purportedly enable them to make vital contributions to some branch of society's economic process. If the demand for these skills goes up rapidly - e.g., if, thanks to new technologies or marketing opportunities in agriculture, housing or health service delivery, employers need an enlarged or differently skilled workforce, or if related opportunities for viable self-employment increase - then so also will demand for the training sequences that provide the skills in question or that open doors to the new employment and self-employment opportunities.

Nonformal education is obviously by no means exempt from these dynamics. Demand for NFE programs is often highest when they are perceived to provide access to other valued opportunities or to furnish skills and knowledge that create new means of viable self-employment and/or materially increase the likelihood of desirable employment. Where supply of such training opportunities is subsidized, costs to participants are lower and the level of expected benefits may not need to be as high to attain the equilibrium point at which motivation triggers participation, but those costs still include the value of time lost and opportunities foregone; and costs to "society", as we shall see, are no less but only distributed differently. Where prospective students and their families bear a larger share of the cost of educational services, the level of perceived benefits will obviously need to be proportionately higher for supply and demand to reach equilibrium and for enrollment and attendance to meet expectations. Nonmonetary costs and unmarketed benefits also obviously enter into the calculation but, from the economist's point of view, do not change the fundamental dynamics of supply and demand.

The market metaphor may seem crass when we are speaking of satisfaction of the "thirst for learning." Much learning is indeed of a self-directed variety that does not enter into market transactions, but

even it may be influenced by forces of supply and demand: arguably people are more likely to try to teach themselves things that promise later to have a social exchange value in addition to the personal satisfactions or utility that these lessons bring them. In any case, given the important socialization role assigned to both formal and nonformal varieties of education in most societies - that is, the function that they serve in preparing and often certifying people for future opportunities - a program that ignores these dynamics does so at its peril, and an evaluator who ignores the forces of supply and demand at work in educational activities risks missing a good part of what is happening.

Notions of efficiency and social welfare

If there is one core notion in the economic analysis of education, it is probably the concept of "efficiency." The term has some specific meanings in economics, plus some more general connotations in popular usage. *Webster's New World Dictionary* defines efficiency as "the ability to produce a desired effect, product, etc. with a minimum of effort, expense or waste" or as "output divided by input." Economists generally distinguish "technical efficiency" from "allocative" or "price" efficiency. The former describes a production process that yields the greatest *quantity* of outputs for a given set of inputs (or uses the least quantity of inputs to produce a given set of outputs). The latter term includes monetary valuation of inputs and outputs. The most "allocatively efficient" production process is therefore the one that produces the greatest *market value* of output relative to the monetary cost of input, or that achieves the greatest "value added."

In the planning literature on education, a distinction is also frequently made between *internal efficiency* and *external efficiency*. "Internal efficiency" can be defined as the ratio of learning outputs to resource inputs and therefore concerns both the economy and the effectiveness with which an educational program produces its immediate (internal) results - gains in skill and knowledge and development of new attitudes. "External efficiency," on the other hand, refers to the ratio between the after-graduation or ultimate *outcomes* of the educational process and its resource costs and thus deals with the domain of job performance, personal satisfactions, and the economic or social impacts of the program (cf. Chapman & Windham 1988).

The purpose of economic evaluation is to improve the efficiency of programs and so of the collective social endeavor that they represent - in other words, to increase the ratio of relevant outputs to necessary inputs. The notion of efficiency itself has, however, at least four significant weaknesses that it is important to keep in mind, given particularly the way in which the term is bandied about, if not brandished, in high-level discussions regarding educational funding. One concerns its dependency on good prior definition of objectives; a second concerns the neglect of distributional issues that usually accompanies it; a third relates to underemphasis of process concerns; and the last has to do with the linear assumptions underlying the concept of efficiency. Let's consider these problems briefly one by one.

Prior definition of objectives

The notion of efficiency really has no content and meaning until objectives are defined, because it simply concerns the maximization of the value or magnitude of objectives that one can achieve with a given set of resources (or, conversely, the reduction of the level of resources required to reach a given level of goal accomplishment). If our objective in a training program is essentially to change people's attitudes concerning village women's aptitudes for continuing education or to create written literature in a local language and these goals are demonstrably accomplished, then the program may be very "efficient" even if none of its beneficiaries sit for the primary school leaving certificate or markedly increase their agricultural productivity. Too often, application of the efficiency criterion means measuring program success by the ratio of the degree of accomplishment of someone *else's* objectives to the resources used by those involved. Logically economic evaluation - or any evaluation heavily reliant on output measures - should be preceded by careful review of how and by whom those objectives are defined, perhaps by the sort of participatory "visioning" described in an earlier chapter. Definition of objectives is more of a political than a technical task.

Neglect of distributional issues

Economists - at least those of the dominant school - tend to be allergic to distributional considerations. Maximizing the ratio of outputs to

inputs says nothing about *who* furnishes the inputs and *who* receives the outputs. Is a program that produces 100 graduates with a given amount of resources but derives its inputs from inequitable taxes on the poor, and which trains in great majority members of the elite class, more socially efficient than one that produces only 50 graduates but generates resources by contributions or levies proportional to wealth and includes in its student body representatives of all social strata?

A simplistic application of the notion of efficiency would lead one to think so. Economists tend to reason purely in terms of aggregate benefit for "society as a whole." Yet no one has ever met "society as a whole," and all known societies are divided into differentially endowed and differentially privileged groups. The question to be asked is "efficiency for what and for whom?"³ Maximizing the ratio of outputs to inputs is important when resources are scarce - as they almost always are - but singular preoccupation with the numbers begs an important part of the question. How interested people are in economizing resources depends in good part on how they think the savings will be used and who will decide the question.

Omission of process

Though cost-benefit ratios may highlight areas of efficiency or inefficiency, they do not explain *why* these situations prevail. In order to improve things, one needs to know what explains the meager results or high consumption of resources that characterize current project implementation; and such insight can only come from careful investigation of the *process* by which inputs are transformed into outputs - that is, in educational programs, the methods and activities by which learning is promoted, instructional sessions are organized, and applications of new knowledge are encouraged.

Overemphasis on efficiency criteria may serve to obscure these critical topics and leave evaluators with insufficient amounts of the (often qualitative) data necessary to make a good diagnosis and arrive at recommendations for program improvement. In short, the concept of efficiency is afflicted with the "black box" problem described earlier. If we concentrate strictly on what goes in and what comes out, we lose sight of *how* the first is transformed into the second and what might be required to render the transformation more effective or satisfying. Box 13.1 illustrates this point.

*Box 13.1. Qualitative benefit-cost evaluation in Mali:
"Rentabilité" and "rentabilisation"*

Funders of the evaluation of functional literacy carried out in western Mali in 1976-1977 were very eager to have data on the real "rentabilité" of the program - French for its "profitability", its yield as an investment or rate of return. In other words, how did quantifiable individual and social benefits compare to resources consumed? After a first assessment of program processes and results, evaluators responded with some data on unit costs and effects. They pointed out, however, that a more important issue seemed to be the "rentabilisation" of the program - a word which means efforts and investments undertaken to *make* an enterprise profitable.

Impacts of the literacy program in western Mali were in fact quite circumscribed, but not so much because people hadn't or couldn't learn the new "technology" of writing in western script as because there were as yet few naturally-occurring uses for literate competence in the Bambara language or for increased numerate skills, and the most important potential in this regard - turning over agricultural marketing and credit operations to villages with sufficient numbers of literates - remained unrealized due to the unwillingness of the responsible agencies to divest themselves of these powers. The evaluators' point was that the one element of the program needed to make literacy profitable in western Mali had not been provided and, in the words of the children's rhyme, "for want of a nail ... the kingdom was (being) lost." Under such circumstances, it seemed as important to perform a qualitative cost-benefit evaluation which identified the *blockages* impeding greater return as it was to calculate rates of return to the last decimal.

Linear assumptions about development

Public administration researchers Caiden and Wildavsky (1977) have suggested that efficiency in fact depends in part on having reserves and fallback options that make output sufficiently reliable - and potential savings sufficiently large - to give a reason for rationalizing production and cutting cost. Resource-poor organizations or societies lacking in these reserves are typically beset by conditions of severe uncertainty that make efficiency efforts much less relevant and attractive. In agricultural economics, there has been for thirty years an ongoing debate about the "efficient but poor" hypothesis, mainly

centered on whether subsistence farmers with few resources have any reason or means to be "efficient" in their allocation of them.

The question, in short, is whether efficiency has something to do with the absolute *level* of resources as well as with the ratio of inputs to outputs. As resources fall toward some subsistence level, effects may no longer be linear. So many needs have to be met simultaneously to keep the ship afloat that the production function metaphor of economics with its single output and *ceteris paribus* ("all other things being equal") assumption is no longer valid. Resources are not "scarce," in the economic sense of insufficient to meet a series of well-formulated and potentially feasible alternatives; they are simply threadbare. As a West African proverb succinctly puts the matter, "It takes water in the belly to draw more from the well." The notion of efficiency may therefore not apply overly well to situations of extreme penury where people must figure out how to capitalize their own resolve or modify inequitable distributions of resources.

This point has particular poignancy for the supporters of nonformal and adult education, the eternal "poor cousin" of the educational system. As a participant in a Burkinabè seminar on literacy evaluation said in reaction to a speaker's advocacy of good management training in post-literacy programs a few years back, "One doesn't manage nothing!" ("*On ne gère pas le néant!*") It could similarly be asked whether the paraphernalia of economic analysis makes much sense in programs which are often barely surviving, last funded and first cut among institutional expenditures. But if nonformal education agencies and advocates are - with the exception of some training programs in business and industry and the (increasingly) occasional donor-supported functional literacy effort - typically strapped for resources, that is because resources have gone elsewhere for reasons often explained or at least justified by the sort of economic reasoning that we are discussing. Better therefore to learn to beat the devil at his own game.

The tools of economic evaluation

Economists currently tend to evaluate education programs by assessing and then comparing their costs and benefits. Since the resources

spent on a given educational program could be devoted to an alternative one, to a revised version of the existing one, or to some entirely different dimension of social welfare, they reason that it is important periodically to assess what costs and benefits the execution of a program does in fact entail, as well as the balance of the one over the other for the different groups involved in the program.

This is obviously not simply a subject of idle curiosity for educators. It has a real bearing on decisions regarding future allocations of resources to programs, since those decisions generally must be justified to some extent on the basis of demonstrations of the cost-effectiveness of the activities or the net benefits accruing to key target groups. This is one reason why the economics of education is so frequently and closely associated with educational planning.⁴

Manpower supply and demand approaches

Cost-benefit comparisons are not the first or only economic approach to educational planning. An older tradition, initiated just before and during the Second World War and much in vogue in the 1960s, is manpower supply and demand analysis. Though roundly criticized in subsequent years, it is still in the human resource economist's "bag of tricks."

The premise of manpower supply and demand analysis is quite simple. If we can predict how economic activity in a region or country will evolve over the next twenty to thirty years, or how we would like it to evolve and believe it can, then we should be able to project as well the kinds and quantities of trained people that will be needed to staff these activities. And such estimates give us a basis in turn for projecting desired outputs of the different branches of the educational system, comparing them to current capacity, and identifying the kind of investments in the infrastructure and personnel of relevant training institutions and education programs that would allow us to avoid labor shortages on the one hand and oversupply of the wrong kinds of skills on the other.

The problem, of course, is that things are not quite so simple. Economic forecasting is notoriously unreliable beyond the short term, there is no one-to-one equivalence between kinds of training and types of jobs, employers would often prefer oversupply of qualified person-

nel to labor market equilibrium because excess human capacity drives down wages, and certain kinds of training (like medical education) require such a long lead time that efforts to ensure adequate quantities of graduates in such a distant future chronically miss the mark. As a result of such difficulties, manpower supply and demand analysis fell into some disrepute in the 1970s, but never entirely disappeared. It was often remarked, in fact, that, though the technique had numerous flaws, manpower supply and demand *analysts* themselves turned out to be quite useful people. Whatever our ability to predict the future, it is no less important to have someone around who is paying attention to the critical area of education-job market linkages and is able to analyze trends with some insight. Moreover, though labor demand projections and even educational system capacity analysis may not be an exact science, the practice of this sort of assessment at least allows one to become aware of impending gross imbalances between types of training and skill needs in the economy.

Since the 1970s, therefore, cost-benefit analysis has generally replaced manpower supply and demand approaches as the tool of choice for examining the economic consequences of educational investment. This technique has its strong and weak points as well.

Cost-benefit analysis

If assessing in some manner the costs and benefits of educational programs, and NFE endeavors in particular, makes a good deal of sense, it does not follow that it is necessarily easy to do. In common sense terms, we all know what benefits and costs are: the former refer to funds, or valued goods and services that we receive, the latter to resources that we have to lay out in order to get them. These common sense definitions pose some problems when we try to use them in the economic evaluation of an education program, however. Three principal difficulties arise, and a brief discussion of them will serve to introduce some of the language of cost-benefit analysis, even if it doesn't resolve the problems themselves! The first difficulty has to do with putting a "price" on nonmonetary items, the second with using existing prices of marketed items as a measure of true value, and the third with trying to aggregate all the elements of a cost-benefit analysis into some single set of measures. Confronting these issues will,

moreover, prepare the ground for a brief discussion of some of the mechanics and techniques of cost-benefit analysis.⁵

Pricing nonmonetary items

Many of the outcomes of importance in NFE programs, and many of the resources that have to be devoted to them, do not have any obvious price - that is, they are not bought and sold in markets and so one cannot easily attach a monetary value to them. That is clearly true in the case of many of the personal and social impacts of education, like changed attitudes, cultural enrichment, intergenerational harmony and so forth. It is also true of some of the key "inputs" to programs: for example, political endorsements, a supportive family context, favorable weather - all resources that can be critical to program success. In short, both the "cost" and the "benefit" sides of the economic evaluation equation for education programs contain a mix of items not all of which are easily quantifiable in monetary terms. How do economists propose to handle the frankly nonmonetary ones?

Economists generally distinguish two categories of these and deal with them in two related but distinct ways. First come items that could be marketed or that sooner or later go into a market transaction. Examples include volunteer labor and marketable skills or attributes acquired through education or training. In the former case, one asks what the person in question could have earned in his or her optimum alternate employment had he or she not chosen to volunteer. In the latter, it is a question of treating the "package" of competencies or characteristics conferred by a given type of education as a salable item whose fair price is the average salary that people of such training subsequently command, compared to the average salary of those with similar background attributes who did not receive equivalent education. (We will look at this case more closely below when discussing "internal rate of return" techniques.) In both examples, one tries in effect to simulate how the market would evaluate the worth of the items or characteristics in question if they were in fact directly marketed, or how it eventually will.

The second category of nonmonetary costs and benefits includes items for which there simply is no existing market, like noise abatement, increased self-confidence or enhanced cultural awareness. In this case, economists tend to use what is called a "willingness-to-pay

criterion": in short, they seek to determine what relevant parties would pay to obtain a given advantage or avoid losing it, or what they would pay to stay free of a certain disadvantage or rid themselves of it, if they were offered a reliable opportunity to do so.⁶ For example, residents of a neighborhood might be asked what increase in homeowners' association dues they would accept in order to ensure that runoff from new roads not contaminate their lake, and this would provide a measure of the worth (to them) of certain levels of clean water. Here as well, however, one is essentially trying to simulate the operation of a market, if only a hypothetical one.

In regard to both categories of nonmonetary costs, what is involved is essentially an exercise in gaming out and therefore estimating what market valuation of their worth would yield if there were in fact such a market. Though the reasoning involved is relatively clear, the mechanics for producing such simulations can get quite convoluted.

The perils of using existing market prices

For items that can be bought and sold, whether labor services or material goods, market prices do represent what it actually costs to get them or what one can actually earn by producing them, and that is a very important consideration for any program. Programs meet their budgets or don't depending to a great extent on whether or not they accurately foresee and balance revenues and expenses. But do market prices represent the real "worth" or "value" of the items in question? And what do we mean by "real worth" anyway?

The first question is "worth for whom?" An extra 1000 calories a day may mean little to someone who is consistently well fed but a great deal to someone who is starving. As noted, neoclassical economists, the sort traditionally employed by funding agencies and banks, prefer not to deal with what they call "distributional questions." The essential concern from their point of view is to increase the overall sum total of goods and well-being - the generalized "welfare of society" - and allow market forces or political forces to handle their allocation to different users. This, of course, may well be a formula for the rich getting richer and the poor poorer. NFE advocates will in any case want to note that distributional issues typically have to be raised

explicitly in economic evaluations because they tend to get overlooked. We will return to this point in a moment.

Even supposing that one can talk, as neoclassical economists do, in terms of a general social good, equal to the sum total and average worth of goods and benefits produced in a given region or country, how do we know whether or not - or to what degree - market prices accurately represent the generalized social worth of the items in question? Let's recall the economist's definition of a resource's true cost: namely, the value of what it can produce in its best alternative use. For market prices to express this value, even according to neoclassical economic theory itself, there must be in place a system of exchange that ensures absolutely free circulation of goods and services to all possible alternate uses and guarantees allocation strictly by productivity and performance criteria. In fact, of course, no such "perfect" system exists, and even purportedly "free" markets tend to be rife with price fixing, monopolies, barriers to trade that protect privilege, and a variety of other flaws that make actual prices diverge markedly from whatever "true social value" would be.

Phenomena of this sort are so prevalent that economists have taken to using a method called "shadow pricing" in an effort to compensate for the shortcomings of market data. "Shadow pricing" is a procedure for attempting to determine by simulation (yes, that recourse again!) what the price of some resource would be if markets were in fact fully free.

For example, in some country or region economists may feel that the salaries of university graduates are kept artificially high by custom and accumulated privilege that have little to do with the actual social productivity of these people. They may then attempt to determine the true "efficiency cost" of this resource by shadow pricing - that is, they estimate what the real average productivity of university graduates is in their actual employment situations and how this performance would be valued on the market in the absence of artificial supports of privilege. The corrected figures will then be used in calculations of the benefit-cost ratio for expanding higher education. It is evident by this point, however, that the exercise is getting a little imaginary and that, though its results may be of some assistance in decision-making, a few grams of salt must be administered with them!

The difficulties of aggregating costs and benefits

The third difficulty with the common sense notions of the costs and benefits of a project arises because costs and benefits do not necessarily occur during the same period of time. Typically costs come first; benefits are realized later. They also are not necessarily incurred by the same people. Even presuming we have valid monetary valuations of all relevant resources, how do we roll into a single ratio items with such varied incidence?

Two issues are involved here: aggregating benefit and cost estimates over people and aggregating them over time. The former raises the distributional questions mentioned a moment ago. In NFE these problems cannot be easily ignored, because most such programs have implicit, if not explicit, distributional objectives: one is trying to bring benefits to specific population groups, not simply to the national GNP. The only way to handle this issue directly in cost-benefit analysis is to assign "weights" to the benefits accruing to, or the costs incurred by, different groups. For example, it might be decided that the benefits accruing to village women or the costs falling upon them should be valued or "weighted" twice (or half again) as heavily as those accruing to men, for compensatory equity reasons. Many economists shrink at the "messiness" of this approach, but in fact the process of discussing the weights to be given to different project outcomes and trying to distinguish which costs and benefits accrue to whom is really more important than any magical numbers finally derived from the calculations.

The second concern involves aggregating over *time* - that is, somehow devising a common denominator for costs and benefits incurred at different periods and under different circumstances, some already past and others only projected for the future. A little farther on we will look at the actual operations used for determining the "net present value" of a stream of costs or benefits strung out over time. Here two more fundamental conceptual issues are of interest. First is the basic notion of the time value of money. A sum of money or a set of resources that will be expended or acquired a year from now does not have as much value, financially speaking, as that expended or acquired today. The reason is that resources available today could be invested (in a savings account, for example, or in a commercial venture) and - considerations of risk aside for the moment - would

yield a larger amount in one year's time, an increment equal on the average to current interest rates. In fact the fundamental economic notion underlying educational programs, formal or nonformal, is that of "investment": we are devoting resources to this kind of instructional effort now because we think that it will pay even greater dividends - in terms of cash, personal satisfaction and social well-being, or some combination of them - in the future.

The other notion involved in the time value of resources is that of uncertainty. A bird in the hand, they say, is worth two in the bush. We know what we have today or have had in the past. Whatever the potential or promise, however, we are not sure what we will have a year from now. There is always some risk in banking on projections, and people's estimation of the riskiness of the future, and their tolerance for different levels of risk, may vary considerably. In financial markets, economists generally assume that these predispositions are already factored into interest rates. But in talking about the future benefits of education and training efforts, an explicit effort must often be made to take account of the probabilities and risks involved in expected outcomes, for these are not reflected in interest rates.

Levels of economic analysis

These then are some of the perils in quantification and monetary expression of benefits and costs. The inputs and outputs of NFE programs are much more than monetary, though important subgroups of them can be expressed in quantitative or monetary terms. Economics offers some methodologies for "monetizing" nonmarketed and qualitative factors, but they tend to get a little far-fetched and are not frequently used. For related reasons, Squire and van der Tak make a distinction between three separate levels of cost-benefit analysis of projects: financial, economic and social (Squire and van der Tak 1975).

Financial analysis

The first level, or the "financial analysis," involves the use of actual market prices for monetary costs and benefits in order to examine the

cash flow and present (or likely) profitability of the proposed project. The task is in many respects an accounting one.

Though obviously most applicable to commercial enterprises, this level of analysis does have implications for education projects as well. In a prospective sense, it generally entails analyzing the "fiscal capacity" of the sponsoring organization (government, NGO, community organization or business). Can this entity meet the anticipated flow of costs for the project or - in the case of an ongoing program - continue to cover those that an analysis of trends leads one to expect in the foreseeable future? In a concurrent or retrospective sense, financial analysis involves assessment of the program or organization's financial "performance": that is, how well it has met its recurrent and investment expenditure needs, what kind of balance it has established between income and "out-go," what in effect is its balance sheet or report card on financial management

Economic analysis

"Economic analysis" constitutes the second level in the Squire and van der Tak framework. It involves trying to translate this market data into a valid measure of general social worth by using "efficiency prices" for all social resources involved. In other words, one calculates shadow prices for the whole range of costs and benefits of the project, including non-monetary consumption benefits (like improved health) that must be priced with the aid of a willingness-to-pay criterion.

Social analysis

The third and most inclusive level of analysis is the "social" one. Whereas economic analysis deals only with aggregate net benefit and cost, social analysis incorporates distributional objectives, and "social prices" should reflect the related weights. At this juncture, therefore, decisions must be made about the relative emphasis to give to the costs incurred by, and the benefits accruing to, different groups of stakeholders. These allocational matters cannot be resolved on purely technical grounds.

In this way, a number of the technical issues involved in the conduct of cost-benefit analysis are arranged in a graduated sequence. The problem - and part of the professional mystification of economics - arises

from the fact that the reputation for hard-nosed clarity that adheres to financial analysis and audits in their own particular domain spreads over (or is designedly applied) to the *social* analysis of economists, where judgments are much less certain. As we shall see, the assistance that economic methodology can offer in specifying and scrutinizing the costs and benefits of educational programs - and in setting their operations in a larger context of social impact - should be very welcome to NFE personnel. But one needs at the same time to be on guard against excessive or overly literal application of the economic viewpoint.

Comparing costs and benefits

How then do economists put together information on benefits and costs in order to reason about the worth or productivity of programs, rank them with respect to each other, and present these viewpoints as input to educational planning decisions? The latter part of this chapter will be devoted to briefly reviewing a series of these methods. In all cases, economists approach the issue from a comparative perspective: that is, they do not talk about the absolute worth or value of a program, but of its worth compared to the alternatives. Economics claims to be a method for allocating scarce resources to competing ends, so the question always is, "Are we getting the best results possible from the time, energy and material devoted to this activity, or would we do better to use the same resources in another fashion?"

Least cost analysis

As the name implies, the most elementary type of economic analysis involves quite simply comparing proposed or ongoing projects to each other from the sole perspective of their cost. It is assumed that the alternatives are roughly equivalent in their potential for accomplishment of relevant objectives (or that any significant differences are too difficult to measure), and the prime concern is therefore whether they fit within cost limits or which one(s) can be implemented at the lowest cost.

Though this is an admittedly primitive form of economic analysis, it does have its own data requirements and is frequently used when information on outcomes is sparse and/or time for decision-making is very

reduced. The problems of measuring and pricing benefits are conjured away, but one still needs good aggregatable cost data.

Cost-effectiveness analysis

The next step "up" in the hierarchy of economic evaluation is represented by cost-effectiveness analysis. "Effectiveness" refers to the productivity of a program - how good it is at producing the results it was designed to achieve, or perhaps *any* related positive outcomes. Effectiveness is basically a measure of benefits, and one expressed not in monetary units but in the various units of output: students trained, lessons learned, training centers created, children immunized and so forth. Moreover, one can focus on either the "internal effectiveness" or the "external effectiveness" of an educational program, the former concerning the magnitude of its instructional results and the latter the magnitude of its larger social or economic effects. When effectiveness indicators are compared to cost, one obtains a "cost-effectiveness" ratio (results divided by resources), sometimes expressed as a "unit cost": for example, 8.5 months of effort or \$500 per literate student. This is therefore a measure of what above was called the "technical efficiency" of a program.

One way to jointly analyze the costs and benefits of an educational program is to attempt to calculate the cost per unit of output, much as a manufacturing firm determines the cost of producing each of its final products and compares this with prevailing prices to see if it is worthwhile continuing to produce the items in question. Thus many economic evaluations of literacy programs attempt to calculate the "cost per literate" of the methodology used. The problem in the case of a social service product like education is that it represents several kinds of output and they are generally rather poorly measured. The true immediate output of training is the knowledge or skill acquired, but there is little agreement about how to divide knowledge acquisition into units, or about how to measure and cost the units. Generally, evaluators fall back on proxy measures like cost per student-year and cost per graduate for lack of data and clarity about true learning outcomes. Unit costs (the cost per student in a given program) are generally translated into costs per graduate by calculating the number of student-years required to produce one graduate (total years spent in the system by all students over a given period divided

by the number of graduates in the same period) and then multiplying this by the unit cost.

The problem is that these calculations give no value to what may have been learned by drop-outs and treat graduation (or "literacy") as a standard unit of output. How much of this cost is assignable to math learning, how much to attitude acquisition, and how much to some of the more imponderable but arguably no less important outcomes of education, like self-confidence?

Such cost-effectiveness measures are therefore a bit wooden and bear a very indirect relationship to the learning variables which are of prime interest in an evaluation. As a consequence, to make them useful for decision-making, something more must generally be said about the *quality* of outputs costed in this manner. Such is the strategy adopted, for example, in consumer reports or digests - the magazines devoted to evaluating alternate models of different products available on the market, as an information service to the consumer, who is often bewildered by the claims and counter-claims of publicity. These reports typically talk about the "quality-price ratio" of different products and seek to inform judgment about where one gets the most for one's money, taking plural criteria into consideration. Inexact as such calculations may be, they perform a valued service. And as difficult or tenuous as it may sometimes be to put a price on educational outcomes, it is certainly a step forward to bring cost considerations into the evaluation of education programs and to align them with assessments of benefit. They can scarcely be left out of decision-making.

Full dress cost-benefit analysis

Cost-effectiveness analysis is used when one has monetary data for resources consumed, but where output, though measured, is not expressed in monetary terms. To the degree that *both* the costs and the benefits of a given cycle of education or educational program can be expressed in monetary terms, it is possible to push the economic analogy even further and to consider the form of education in question as an *investment* which yields financial returns over time, much like what happens when one purchases an annuity or puts funds into a business

venture. We are now dealing with what was termed above "allocative efficiency."

This means, of course, that the only advantages and drawbacks of an educational program taken into consideration are those to which a price tag can be attached. Theoretically speaking, we have seen, economists and planners have ways of "pricing" almost any input or outcome associated with education. For those inputs or outcomes (like leisure foregone or cultural awareness gained) not normally bought or sold in markets, techniques exist for determining and measuring "willingness to pay" - that is, the amount that people would be ready to spend in order to obtain (or to keep) the advantage in question, or in order to avoid the disadvantage in question. Practically speaking, however, these sorts of calculations are rarely performed in economic evaluations. Expenditures are habitually restricted to financial outlays plus monetizable "opportunity costs," or the income foregone by virtue of participating in education. Benefits focus on measurable increments in lifetime income associated with the level and type of education in question.

It has also been noted that the monetary (or monetizable) costs and benefits of an education program generally occur at different points in time. Typically, one incurs costs during participation in the educational program, whereas the pecuniary benefits only begin to show up in the form of improved job prospects and increased income several years later. How can these data be converted into a rate of return estimate? Two methods are most prominently cited and used: the net present value approach (which includes rate of return analysis) and the earnings function approach.

Net present value approach

The first and by far the more frequently used approach of the two entails computing the "net present value" (NPV) of the stream of costs incurred and the NPV of the stream of benefits anticipated. Calculating the net present value of a sum of money (cost or benefit) is just the reverse of calculating the future yield of a sum invested today. Both operations involve using an interest rate that represents the current "price" of investment funds to the borrower or, equivalently, the rate at which a sum of money invested today is expected to grow in the future. In determining the future yield of a sum of money invested

today, one multiplies the sum by 1.00 plus the interest rate, this latter expression raised to a power equivalent to the number of years during which the money will be allowed to grow. The reason for the exponential operation is, of course, that the investment is compounding: the earnings in the first year become part of the principal in the second, and so forth. In mathematical notation,

$$\text{Future yield} = \sum_{t=0}^n B_t (1+r)^t$$

where t	=	time period
n	=	number of years during which investment will grow
Σ	=	symbol denoting summation of the following expression from the subscripted time period to the superscripted one (i.e., in this case, from the time period "0" to the last year in the growth period, or the "nth" year).
B	=	amount of benefits in any period of time ("t")
r	=	interest rate

To establish a net present value, one does just the reverse: that is, the future yield (or stream of benefits) is deflated by the interest rate to determine what sum of money would have to be invested now in order to generate such a return. This is the "net present value" of the future sum of money. The formula is as follows:

$$\text{NPV} = \sum_{t=0}^n \frac{B_t}{(1+r)^t}$$

To determine the effective rate of return (ROR) to an educational program, the problem is changed only slightly. In each of the examples above, one sum of money and the interest rate are known and these two variable values are then used to compute the other sum of money. In ROR calculations, we have an estimate of *both* sums of money: the costs to be incurred and the future benefits to be realized. The problem is to use these two variable values to compute the effective interest

rate. By definition, the interest rate that would make a stream of costs ($C_1, C_2, C_3 \dots$) with a net present value of

$$\sum_{t=0}^n \frac{C_t}{(1+r)^t}$$

grow into a stream of subsequent benefits ($B_1, B_2, B_3 \dots$) with net present value of

$$\sum_{t=0}^n \frac{B^t}{(1+r)^t}$$

is precisely the one that would cause the two net present value figures to be equal to each other. The rate of return is therefore determined by solving the following equation for "r":

$$\sum_{t=0}^n \frac{B_t}{(1+r)^t} = \sum_{t=0}^n \frac{C_t}{(1+r)^t}$$

In practice, the problem is handled by repeated hit-or-miss calculation until one lands on the rate that causes the two sums to be equal. The process is happily made much quicker and infinitely less painful by the computer: most spreadsheet programs can do these operations easily and have built-in routines for internal rates of return. One need only enter the series of benefit and cost data. Perhaps the spread of such software explains in part the increased use of rate of return calculations!

The advantages and disadvantages

The advantages of this approach are that it is relatively simple and straightforward and that it provides a convenient means for handling costs and benefits occurring over a period of several years at different points in time and for collapsing them into a single expression. The

NPV approach also has some marked disadvantages, however, of which three are worth mentioning here. First, the method does not work well when benefits and costs are not segregated into discrete periods of time - when, for example, the person realizes some benefits for a number of years, then incurs again some educational costs. Second, nothing in the procedure allows one to estimate to what degree the "benefits" realized are really attributable to the education undergone. Third, the data used to estimate rates of return by this procedure are almost unavoidably cross-sectional - that is, they represent with greater or lesser degrees of accuracy what people possessing the degrees in question are *presently* earning on the labor market, not what they will earn during the working life of students currently contemplating continued education. Such longitudinal data are seldom available.

The resulting statistic is designed to be interpreted in much the same manner as an interest rate on savings or a level of financial yield on business investment. In other words, if the figure is above current rates of return on capital elsewhere in the economy, then the type of education in question would appear to be a good investment; if it is below, then the investment looks less attractive.

But a good or bad investment *for whom*? In partial response to this question, economists and education planners differentiate the *private rate of return* to educational expenditures (PROR) from the *social rate of return* (SROR). In calculating the former, one counts only the costs of education actually incurred by individual students (including nonetheless the worth of their forfeited time), and income benefits are restricted to after-tax earnings. The PROR thus expresses how profitable a given cycle of program of education appears to the prospective student him- or herself.

In calculating the social rate of return, on the other hand, *all* costs are included, since they represent social resources that can no longer be put to alternative use, and the income measure corresponds to before-tax earnings, since the entire value of the individual's work accrues to society, even if a part of it is taxed away and is no longer in the worker's possession.

Though private rates of return provide a means for analyzing and predicting the level of popular demand for different kinds of education, it is evident that the social rate of return has greater potential policy relevance. A technique that allows one to compute the social

worth of educational programs from available data and to compare their investment potential to that of alternate targets of government or NGO funding like infrastructure projects or farm price support policies has inherent attractions for decision-makers. Unfortunately, the SROR is also the more tenuous of the two measures. The principal problem arises from the built-in assumption that income is an accurate reflection of the social worth of the associated work. In neo-classical economic theory, this assumption holds under conditions of perfect competition when all markets are in equilibrium. Such conditions are never realized, however, and market "imperfections" introduce varying levels of distortion into the relationship of income to productivity. The link becomes even weaker when cross-sectional data on current levels of income are used - as they must be - to project future income differentials between, say, primary and secondary school graduates and to infer actual productivity differentials between them.

For this reason, and for others illustrated in Box 13.2, social rate of return figures must be interpreted with considerable caution. They highlight an important issue and provide a useful language for addressing it, but it is as important to read the assumptions as the conclusions. And that is too seldom done.

Earnings function approach

The second major technique for determining rates of return - the "earning functions" approach - permits one to differentiate returns to schooling from returns to professional experience, but it has a number of other disadvantages and is not very frequently used in actual policy research. Data on the level of schooling and years of professional experience of a sample of workers are regressed on the natural logarithm of their earnings.

In this formulation, the coefficient of the schooling variable represents the average marginal productivity of education (i.e., the relative change in earnings per unit change in education) and is therefore equivalent to a rate of return to education net of the effects of professional experience. However, it is not easy to incorporate information on the cost of education in this model and it therefore is of little help in estimating *social* rates of return. It is thus very infrequently used for policy purposes.

*Box 13.2. Rate of return analysis:
In Haiti and Botswana it all depends...*

Rate of return analysis became one of the mainstays of the educational sector assessments conducted by the World Bank and agencies following the Bank's lead in the 1980s. During some of the assessments performed under the aegis of the IEES (Incentives for Improving Basic Education) Project, an USAID-funded initiative to support better educational planning and research in developing countries, the indeterminacy of the method began to show through. Like much in economics, it all depends on the assumptions.

The general role assigned the economic analysis in these sector assessments was to demonstrate the profitability of basic formal education and so to support the related conclusions which most such studies reached. In Haiti, though, evaluators took the methodology seriously and generated over fifty different rates of return for the myriad species of primary and secondary education in Haiti, with returns varying from nearly nil to respectable levels depending on where one was in the social hierarchy - meanwhile admitting that it was all essentially a heuristic simulation, because reliable data on actual incomes of primary and secondary school graduates in the country were simply not available. In Botswana, initial calculations suggested a poor return at the time to primary education, but the economists on the sector assessment team, mindful of the political target to be reached, changed the assumptions of the analysis and managed to raise the rates to an acceptable level.

This slipping and sliding is not atypical of the calculation of rates of return, which depends heavily on the assumptions made, and generally on secondary - if not on simulated - data as well. The exercise of assembling and examining information on the costs and benefits of different forms of education is certainly an extremely important one, but the distillation of this data into a single index or set of indices to charm decision-makers poses a multitude of problems. The best defense for NFE evaluators at this point may be a good offense: value the data, know the technique, and vigorously question its validity in specific uses.

Applications to nonformal education

What then is the importance of these sometimes arcane notions and techniques of economic evaluation for nonformal education programs and for their evaluators? Two considerations stand out. First, whatever the weak-

nesses of economic methodology when applied to NFE programs, economics is the language of much educational planning and decision-making, particularly in Ministries of Finance and the policy units of donor organizations. If only for reasons of self-defense and ability to understand discourse that may determine the fate of such programs, evaluators do well to become familiar with the terminology and ways of thinking of educational economists.

The second consideration is more positive. The basic notions of economic evaluation contain some partial truths of great importance for NFE programs - and in a world of plentiful falsehoods, a genuine partial truth is certainly worth using and keeping, as long as one doesn't take it for holy writ. Recognizing the resource *costs* of training programs and trying to specify their *benefits* in concrete, if not always monetary, terms are critical exercises for the organizers and sponsors of NFE, particularly in an epoch when such programs are increasingly called upon to justify their existence or to finance themselves.

Arguably, in fact, assessing post-training outcomes in terms that can be compared to cumulative resources invested on either a cost-effectiveness or a cost-benefit basis is strategically even more important in NFE than it is in traditional formal education for a very simple reason. Formal education generally enjoys a level of cultural and institutional justification that makes such programs - in many people's eyes, at least - their own excuse for being, whereas nonformal education is typically required to prove its merits by demonstrations of personal and social impact. With some hyperbole, one might say that in NFE the tree is known by its fruits, in formal schooling by its pedigree. It follows that NFE evaluators and organizers need to be particularly adept at understanding and expressing the "downstream" consequences of training and comparing them to the effort these programs require, both in order to make convincing analyses of their worth and to provide information that will help program designers better adapt their offerings to social needs.

Economic evaluation provides a language, a methodology and a perspective to help in that endeavor. Even in the case, for example, of those important outcomes of NFE that cannot, and perhaps should not, be expressed in monetary terms, the attempt to define them, to gauge their magnitude and to ask if they are worth the sacrifices required can be highly edifying. Such an exercise should not be considered the exclusive province of those trained in economics, though the right sort of econo-

mists can be extremely helpful. It is no exaggeration to say that economic evaluation *cannot* be validly performed without a large measure of participation from program staff and beneficiaries.

Campen (1986) has pointed out that the crux of economic assessment lies finally in the determination of "willingness to pay" and in the choice of weights to assign to different outcomes and to different types of cost incidence. The work of identifying and ranking outcomes, of deciding how the relative burden of costs on different stakeholders should be assessed and of drawing conclusions from various combinations of these kinds of data is something that virtually *requires* open debate and the confrontation of differing points of view. Campen himself asserts that, if properly executed, cost-benefit analysis is an excellent vehicle for introducing greater popular participation into program planning while instituting greater accountability. It is a tool, he would say, for empowerment. But it is obviously not one if the procedures and the language totally mystify the people who should be using them. It is the evaluator's task to dissipate this mystery (some of which, it must be said, is designedly created and sustained by economists themselves) and to turn the economic elements of program evaluation into an occasion for better and broader understanding of how resources are being used to instructional ends and how this compares to the palpable impact they are having on people's lives.⁷

Conclusions: Economic metaphor and educational process

Most of the concepts discussed in the foregoing pages spring from the application of a "production model" to education. The operation of educational programs is compared to a sort of industrial process in which inputs derived from the environment are transformed into outputs that are marketed and subsequently returned to the environment, producing a variety of impacts upon it. This metaphor lends itself to quantification and to some useful kinds of analysis, but it has definite limits.

To borrow systems terminology, the production model does not handle plural outputs or multiple criteria very well. It works best and yields quantitative results most reliably when the factors taken into consideration are greatly simplified - as in "student-years per gradu-

ate," "dollars per literate," or "income increments associated with further training." Quantitative models tend to bog down or become uninterpretable when too many variables are entered into the equation. As a consequence, they typically leave out a number of the subtler processes of education - like transmission of cultural values and formation of attitudes - that determine its quality and play a major role in producing its results. Use of any of the derivatives of the production model needs to be supplemented by qualitative analysis and tempered with an awareness that these concepts are largely metaphor and that most of the indicators are only proxy measures of the topics that really interest us.

In fact, when examined closely, the economic calculation of efficiency or of internal rates of return to education is more than a little specious. Like economic evaluation as a whole, it is more useful in a figurative sense as an exercise in aligning and discussing the costs and benefits of training and education - certainly a worthwhile endeavor that is too seldom undertaken in systematic fashion - than as a literally valid means for determining the "worth" of programs. Unfortunately, economists have a tendency to forget this, which is why NFE proponents and evaluators will do well to learn some of the language and analysis of economic assessment, if only in self-defense!

Notes

¹ Of recent years, economists have taken to calling this method "benefit-cost analysis," since benefits are in the numerator of the equation and seem to strike a cheerier note when used as the first word in the expression. However, since the term "cost-benefit" is still better known in everyday parlance and easier to remember because of its parallel with the names of two other prominent kinds of economic assessment of projects - "cost-effectiveness" and "least cost" analysis - it is the one that will be used in this chapter.

² Of course, some learning is entirely self-directed and does not enter directly into exchange. In fact, it can be argued that *all* valid and lasting learning has a self-directed component. That does not invalidate, however, the exchange relations that also characterize much of it - a point discussed further on.

³ This line of argument is much more fully developed in Easton & Klees (1990).

⁴ Textbooks and expositions on the economics of education that will prove useful for evaluators of NFE include Ahmed (1975), Blaug (1972), Educational Testing Service (1979), Johnes (1993) and Levin (1983).

⁵ Further detail on techniques of cost-benefit analysis may be found in Mishan (1988) and Nas (1996).

⁶ The word "reliable" is important in this phrase, because the probability that such offers will really deliver expected results - or, inversely, the *risk* (another key term) that they won't - obviously enters into the calculation, as do the different levels of tolerance of (or taste for) risk of the people involved. We will return to this point in a few pages.

⁷ In fact, as Campen's remarks and Box 13.1 in this chapter -- as well as the study portrayed in Box 9.2 in an earlier chapter -- all strongly suggest, there is every reason to envisage and carry out qualitative and participatory varieties of economic research and evaluation, a brand of study that would merit at least a chapter of its own, not to speak of more attention in academic media. There is much that transpires in or determines the nature of economic transactions that cannot be adequately captured through numbers alone. The field of economic anthropology pays heed to these issues (cf. e.g., Gregory 1989; Plattner 1989).

References

- Ahmed, M. 1975. *The Economics of Nonformal Education: Resources, Costs and Benefits*. New York: Praeger.
- Becker, G.S. 1964. *The Economic Approach to Human Behavior*. Chicago: University of Chicago Press.
- Blaug, M. 1972. *An Introduction to the Economics of Education*. New York: Penguin.
- Caiden, N. and A. Wildavsky. 1980. *Planning and Budgeting in Poor Countries*. New Brunswick, NJ: Transaction Books.
- Campen, J.T. 1986. *Benefit, Cost and Beyond: The Political Economy of Benefit-Cost Analysis*. Cambridge, MA: Ballinger Publishing Company.
- Chapman, D.W. and D.M. Windham. 1988. *The Evaluation of Efficiency in Educational Development Activities*. Tallahassee, FL:

Improving the Efficiency of Educational Systems (IEES) Project (Florida State University).

Easton, P.A. and S. Klees. 1990. "Education and the Economy: Considering Alternate Perspectives." *Prospects*. XX (4): 413-428.

Educational Testing Service. 1979. *A Manual for the Analysis of Costs and Outcomes in Nonformal Education*. Princeton, NJ: International Office, Educational Testing Service.

Gregory, C.A. 1989. *Observing the Economy*. New York: Routledge.

Johnes, G. 1993. *The Economics of Education*. New York: St. Martin's Press.

Levin, H.M. 1983. *Cost-Effectiveness: A Primer*. Beverly Hills, CA: Sage Publications.

Mishan, E.J. 1988. *Cost Benefit Analysis: An Informal Introduction*. 4th ed. London: Unwin Hyman.

Nas, T.F. 1996. *Cost-Benefit Analysis: Theory and Application*. Thousand Oaks, CA: Sage Publications

Plattner, S. (Ed.) 1989. *Economic Anthropology*. Palo Alto, CA: Stanford University Press.

Robbins, L. 1963. *Politics and Economics: Papers in Political Economy*. New York: St. Martin's Press.

Samuelson, P.A. and J. Nordhaus. 1985. *Economics*. 12th ed. New York: McGraw-Hill.

Squire, L. and H.G. van der Tak. 1975. *Economic Analysis of Projects*. Baltimore: The Johns Hopkins University Press (for the World Bank).

Part V
Conclusions

CHAPTER 14

THE CONTINUING EDUCATION CHALLENGE IN EVALUATION

The objective set at the beginning of this book was to review a certain number of topics, both technical and organizational, that might be thought to constitute subject matter for "intermediate" level training of the evaluation personnel of adult and nonformal education programs. I have tried to focus on aspects of the evaluation trade that lie beyond the basics to which practitioners are initiated in training sessions or which they encounter in available introductory texts - intermediate-level lessons which they need to master in order both to strengthen their own competence and to become more effective in training program staff and beneficiaries in evaluation perspectives.

Taking the bull by the horns

Two assumptions underlie this approach. The first is that evaluation requires a combination of skills and methods which few practitioners bring ready-made to the field and that, given the relative dearth of advanced training programs in evaluation methodology (or directly relevant to it) in existing institutions of higher education - particularly in developing nations - practicing or would-be evaluators need to take their "continuing professional education" into their own hands. The second is that good evaluation practice necessarily means involving others in evaluation and therefore developing one's skills and aptitudes for participatory process. The "continuing education challenge" of evaluation is both to ensure the ongoing professional education of evaluators and continually to transform evaluation into an educative and participatory exercise for all involved.¹

This dual challenge of continuing education is of high importance for evaluation personnel for several different reasons:

- Evaluation is a growing field, in both a political and a technical sense. "Politically" speaking, though the era of large centralized so-

cial intervention programs and abundant national funding for education projects appears to be past, movements for greater fiscal austerity and for accountability in social institutions place, if anything, greater emphasis on the importance of evaluating results and using insights gained from experience to render interventions more cost-effective. As a Hausa proverb of recent origin succinctly phrases the matter of accountability, "If nothing has been forged, what happened to the charcoal?"

- In technical terms, the tools and procedures available for program evaluation have multiplied over the last thirty years, and the accumulation of knowledge and technique in the field has been such that practitioners have a wealth of precedents and "clinical experience" on which they can draw - or should learn to draw.
- Like most professional fields, evaluation is at the confluence of a series of disciplines and technical skill areas. Among them are testing and measurement, social science research, policy studies, organizational science, ethnographic methods, and fiscal management. Advances in all these fields may have a bearing on the conduct of evaluation, and no one is inherently skilled in all or able to keep abreast of their evolution simultaneously.
- If, as has been emphasized in this book, evaluation is quintessentially an experiment in "social learning," then evaluators obviously need to practice what they preach. They do this by continuing themselves to be active learners and by helping others to learn as well.
- Improved technique and broadened participation are intimately involved in evaluation practice. On the one hand, evaluators need enhanced technical skills if they are to facilitate truly participatory evaluations and ensure a learning experience of worth to those who take part. On the other, improved technique means and produces little if it is not grounded in the sort of validity that only broad stakeholder participation can ensure for evaluation efforts.

How much, realistically speaking, can one do to meet these challenges? Evaluators regularly find themselves in situations where the pressures of the job and the lack of support and resources for continuing education - not to speak of frequent lack of recognition of what they do as a valid professional activity - make it very difficult to keep up to date in

their field or to envisage further training. Under these circumstances how do evaluators go about both providing for their own "continuing professional education" and involving broader ranges of stakeholders in the learning experience? What follows are a brief discussion of this issue and a few practical suggestions.

Obstacles to enhancing technique and participation

First, it seems advisable to take full account of the difficulties. In my own experience they come down to time shortages, resources inadequacies, lack of social recognition, and widespread skepticism about participatory approaches.

Time constraints

Time is a major obstacle to continuing one's professional education, not just because evaluators are typically very busy with their assessment duties, but because they often "wear several hats" at once. Organizations that cannot pay for full-time evaluation personnel, that do not give this issue much priority or that judge it better to combine evaluation work with other responsibilities regularly assign staff to evaluation functions on a part-time basis. This lack of endorsement can also have the effect of promoting increased turnover among evaluators, a further constraint on the time resource for continuing training, because people whose tenure in a professional position is expected (by themselves or by others) to be short are less likely to invest effort in training and because skill improvement in evaluation, as in most professional fields, is in large part a question of the accumulation of experience.

Resource constraints

Lack of resources impinges in several ways. Evaluation functions themselves are often low-priority and meagerly funded ones within social and educational agencies, except when outside-financed projects include them as prerequisites. Why pay for evaluation if there isn't enough money to pay for the program itself? As mentioned earlier, the most frequent conclusion of the generally well-funded evaluation efforts of the late

1960s in the United States was that nothing could be concluded because the program in question had never been implemented as planned. Evaluation is obviously at best a *second* priority. Administrators rightly feel that it is more important to get something going in the field before worrying about just how to gauge its effects or assess its methodology.

If there are scant funds to underwrite actual evaluation studies, how likely is it that there will be resources to support continuing professional education or advanced training for evaluators, even if a number of "human capital"-type arguments could be made for such investment? The most likely source of support for this type of personnel upgrading has generally been participant training funds from major donors which have the latitude and the means to engage in long-term thinking about institutional development and training needs. But those kinds of projects seem increasingly scarce.

Lack of resources can also refer to the relative unavailability of training offerings, advanced education options and self-instructional technical materials on evaluation topics. Such learning tools are only easily available in a few industrialized countries. In recent years some bilateral and multilateral donors and NGOs have made a concerted effort to provide increased written material of this nature, but resources are still thin on the ground.

The low status of evaluators

Lack of social recognition is a third obstacle in the way of more support for the continuing professional education of evaluators. Evaluation is not widely recognized as a profession or a respectable technical skill area, both because it is a relatively new specialty first largely practiced under the aegis of large donor organizations, and because evaluators are widely seen as "nay-sayers": critics whose task it is to bear bad tidings. In cultures where there is a considerable premium on consensus, evaluators are often consciously or unconsciously seen as dysfunctional elements. (This is why Bhola [1990: 72-73] speaks so prominently of the importance of creating a "culture of evaluation" - that is, a social and institutional atmosphere in which information on program processes and results is regularly generated and widely circulated, and where assessment is welcomed as an opportunity for open discussion and program improvement, not as a form of last judgment.)²

In any case, in the words of the comedian, it is often true of evaluators that they "just can't get no respect." The low status and priority of evaluation work creates an additional *human* resource problem: high turnover among evaluation personnel and difficulty in recruiting people with strong general qualifications into the profession. This in turn makes continuing professional training of evaluation personnel both less likely and potentially less fruitful: people with little personal investment in the field have correspondingly little interest in devoting to it the time and effort necessary to upgrade their competencies, and those with weak general credentials may be poorer candidates for advanced training.

Obstacles to participatory methods

Skepticism about participatory methods and empowering approaches to evaluation is the last hurdle to consider. Despite increasing interest in methods of participatory research and evaluation and the current *zeitgeist* of democratization, participatory approaches are much more often endorsed than implemented, particularly in technical domains and in those that bear immediately on issues of funding. Evaluation falls in both such categories, so some difficulties are to be expected and are usually encountered when one advocates opening up the evaluation and decision-making process to wider participation. On the one hand, technicians do not like the implication that their realm of expertise is easily accessible to amateurs; and, on the other, management staff are not generally fond of sharing decision-making responsibility or creating avenues for increased accountability.

Hope for enhancing technique and participation

Given this array of obstacles, it is scarcely surprising that good continuing professional education opportunities and practices for evaluators, and for those in the nonformal education field in particular, tend at present to be few and far between, or that relatively little is done to spread evaluation competence more widely. The picture is not entirely bleak, however, and it is important to mention some of the reasonings and resources on which one can draw in the attempt to improve it.

The first is simply that evaluation and related planning and policy activities are functions that are not about to disappear, so we are entitled to think and scheme for the long haul. As mentioned earlier, while fiscal austerity in the public sphere has cut down the scope of centrally-funded social and educational programs, it has also increased the premium put on "accountability," institutional "transparency" and different forms of decentralization and democratization - all movements which suppose and require good information about the processes and results of programs, be they central, regional or local in scope, and private or public in nature. It is significant that even while large social programs have passed into relative disrepute, "management information systems" - which are essentially institution-specific evaluation schemes - have become all the rage. One could make a decent argument that increased democratization in fact supposes improved evaluation, because if the "people" are to decide - whether in their capacity of citizens, association members, employees or stockholders - they need information on the basis of which to do so. Evaluation will have its ups and downs and could itself stand to be "decentralized" (which is the point of Chapter Three on participatory methodologies), but it seems destined to be around for quite a while in some form. Or, as the poet Robert Frost once put it, "I never stopped believing in something simply because it ceased to be true. Believe in it long enough and it becomes true again. Most of the news is simply made by truth going in and out of fashion."

Second, it is worth noting that all professions have had some difficulty in establishing themselves and finding a stable role in society. This is typically a process that goes on over decades. Traditional professions (law, medicine, architecture) are thought to require an identifiable body of knowledge, an institutionalized form of training, and functional rules of exclusion - that is, ways to say who is and who isn't a "legitimate" practitioner - in order to control quality and/or limit entry. New-style professions of the "reflective practitioner" type described by Donald Schon (1982) arguably proceed more by *inclusion* than by exclusion - that is, the professional acts as a facilitator and teacher helping other people to learn the basics of the trade. Evaluation is potentially in the forefront of this second wave of professionalism because, as suggested at several points in this text, effective evaluation supposes client participation.³ Moreover, the habits which it promotes - like joint collection and analysis of data and open debate about objectives - are among the keys to

greater "client" involvement in a series of other professional fields as well. So in this respect too there is a bright side to the picture.

Third, note that the knowledge base of evaluation is real though, like any profession, composite and composed partly of precedent. Professions in general lie at the confluence of several different "disciplines" and areas of technical skill. Medical personnel, for example, draw on biology, psychology and sociology in their work, as well as upon laboratory technology, counseling and social work methods and administrative procedures. Their training tends to embrace all these to some degree and to rely in good part as well on the accumulation of clinical evidence from cases where practitioners have had to put all these perspectives together in resolving medical problems. For educational evaluators, the sources include educational psychology, sociology, anthropology, economics, political science and a series of technical domains like inquiry design, psychometrics, administrative assessment and cost-benefit analysis. Precedent derives from the recorded or recounted experience of previous evaluations under varied circumstances as well as longitudinal perspective on their effects. As this book hopefully helps to demonstrate, there is definitely material here for intermediate and advanced study, though it is quite related both to what good educational researchers and administrators should learn in their own higher training and to what is purveyed in policy and evaluation studies in other professional fields like social work, business, and agricultural extension. In any case, this combination of competencies will only be of growing importance to the conduct of educational programs and provides one sort of important "anchor" for the evaluation field.

Finally, whatever its ups and downs, the drive for greater citizen control of - and input into - social services is not likely permanently to reverse itself. Participation as an issue seems here to stay, in one form or another; and evaluation, as argued earlier in this book, provides one of the best "entering wedges" for broadened participation of staff and beneficiaries in organizational affairs. Evaluators who see themselves as the "midwives" or facilitators of good evaluation practice by broader groups of people, as opposed to the sole guardians of true methodology in the domain, will doubtless be the best equipped to share in and support this movement. And organizations that make a practice of assessing their own performance, recognizing and learning from their successes and failures,

and involving both staff and stakeholders in this evaluative effort will arguably have the best chances of prospering in a new century:

Strategies for enhancing technique and participation

Let's turn in conclusion to some of the strategies that evaluation practitioners can use, given both the obstacles and the potentials just evoked, in order to promote their own continuing professional education and to develop their "teaching" and facilitation role. Note first that there are a variety of ways of pursuing the continuing education goal, including self-instruction through study of relevant texts, apprenticeship with experienced practitioners, short-term training workshops or conferences, and different types of renewed higher education (distance learning, certificated experience, residential classwork...). This is another domain where one probably does best to "diversify the portfolio" - that is, to pursue the twin goal of constantly increasing knowledge and of sharing it more widely through a set of varied and complementary means. The critical common denominator is that the strategy employed offer you an opportunity to reflect on your own experience, try out new approaches in your professional practice, draw lessons from them, and involve broader ranges of program staff and beneficiaries in this learning-by-doing. Simply reading texts or even attending classes or workshops will definitely not do the trick. Evaluation is a field of praxis, and knowledge gained regarding its methods and techniques is to be shared.

One principle or strategy to remember is that there is strength in numbers. This doesn't necessarily mean trying to recruit a cast of thousands into one's own evaluation unit. Rather, it means simply realizing that there are probably a lot of evaluators around. Those who work in the field of education have in fact a great deal in common with those working in other fields like public administration, agriculture, health, urban planning and so forth. What's more, impact evaluations of educational programs inevitably involve economic, agricultural, health and other social change considerations with which these people may be much more conversant than are educators; so there is an additional payoff to opening lines of communication and seeking avenues for collaboration. One of the best of these can be simply to create some form of association among people working in evaluation, whether in public or private positions and

whatever the social sector involved, for exchange of ideas, consideration of common problems or needs, and promotion of relevant training opportunities. Professional associations have for a long time been one of the prime sponsors and sources of continuing professional education, generally through short-term training sessions and supply of instructional materials, but also by their efforts to consolidate legitimate formal training sequences in institutions of higher education. In my own home town - a state (regional) capital - an association of evaluators formed less than 10 years ago now draws to its annual conference hundreds of people from across sectors and public-private sphere distinctions.

Within their own organizations, evaluators need to promote the principle - equally valid for all technical personnel - of human resource development "sabbaticals": that is, periods of time that can be taken off from work at some relatively regular interval for more intensive forms of continuing education. Theoretically, something like this is an imperative in all organizations. As Peter Senge argues, the pace of technological and social change is such now that agencies and enterprises wishing to keep up with their environment and ride out daily events need to become "learning organizations" and promote both regular retraining of their staff and a pronounced learning environment in the workplace (Senge 1992). In reality, of course, meager staff and material resources mean that such organizations can rarely fully fund time-off and external advanced education for their employees unless such opportunities are made available by donor bodies.

Still, half a loaf is better than none, and the basic point to get across and hopefully implant in the culture of one's organization is that productive work cannot be maintained without some form of "downtime" for refreshing skills and increasing knowledge - whether it is one day a month, a week or two a year, or a year every seven or eight. Furthermore, one can often devise strategies for what might be called "sweat equity sabbaticals"⁴ - that is, ways to reorganize work among members of a unit or team so that one of them can be "broken free" for advanced training during a specified amount of time, on the condition that sooner or later everyone share in this opportunity. This might be thought of as simply the application to evaluation practice of the "tontine" mechanism so widely used by neighborhood and women's groups in many countries: we all contribute to assembling a little capital and award it to each member in turn.

There are negative consequences to avoid here - like the risk of encouraging layoffs by demonstrating that the same work can be done by fewer people - but the underlying mechanism is worth considering. Paradoxically yet significantly, other staff are most likely to look favorably on continuing professional education for evaluators if evaluation personnel have been sharing their knowledge and promoting participatory practice, because people are more likely to want their tutors or coaches to continue learning than they are to wish to add to the knowledge and privilege of distant technicians. In fact, participatory initiatives that mobilize broad groups of staff and program beneficiaries can also be good vehicles for the continuing professional education of non-evaluation staff, because - if properly handled - they often create the momentum necessary to bring in outside resource personnel who can help to turn the exercise into an excellent opportunity for on-the-job training.

When it comes to continuing training of the formal variety in institutions of higher education, the lack of officially-constituted degree programs in evaluation need not be a major obstacle. There is a considerable debate in Europe and the United States concerning whether or not advanced degrees should be conferred in evaluation in the first place, or whether practitioners should instead take a degree in one of the component disciplinary or technical areas (e.g., sociology, educational psychology, economics of education, research methods, regional planning) and give it an evaluation twist or imprint. In any case, evaluation is a composite field, and the important thing with regard to higher education experience is to undertake study in a closely related domain (like those mentioned) and/or to enroll in a program with enough latitude to allow the student to put together a menu of studies that covers important foundations of the field - and then to apply this knowledge to evaluation practice. To a considerable extent, an evaluation degree is in the eye of the beholder: if you have acquired relevant skills and given your work an evaluation focus, then you are entitled to consider yourself a university-trained evaluator! Your degree must be real, but is then largely what you choose to call it. "Spin doctoring" is nothing new in professional circles. Or as humor in the extension tradition puts it: "Last year I couldn't spell 'agronomist' and now I *are* one!"

Finally, access to relevant and recent written material is a key factor in continuing professional education of whatever type, and typically a sticky wicket in environments where literature of general interest is so

expensive and/or hard to come by, let alone specialized and technical materials like evaluation texts. International organizations can and have done something to alleviate the bottleneck (as witness publications in this series from UIE and DSE), but the likelihood is still rather slim that individual evaluators will be able to assemble personal and up-to-date libraries of internationally known works on evaluation or to subscribe to many current journals in the field. The following recourse strategies are worth mentioning:

- First, without wishing to encourage violation of copyright, one should certainly recall that photocopying provides an alternate way of circulating highly relevant technical materials in places where the originals are unavailable or unaffordable. Jerry Rubin, the late unrepentant "Yippie" of the 1960s and 1970s in the United States, once published and marketed through retail dealers a work simply entitled *Steal This Book!* The one you hold in your hand might be entitled *Photocopy This Text*.
- Second, associations or groups of people can have much better success at assembling resource collections at reasonable cost than individuals - yet another reason for reinforcing some sort of network or association of evaluators. Not only do such groups provide a medium for pooling resources, but it should be increasingly possible to promote direct linkages between evaluation associations in the South and similar institutions in the North along which relevant technical material can be passed.
- The images of linkages and networking just evoked immediately make one think of the possibilities opened up (or, in many countries, still gradually opening up) thanks to the Internet and the World Wide Web - the international computer network that can immensely facilitate both communications among evaluators and access to literature and data bases. Once again, though individual hook-ups with the Internet may be problematic for some time to come in a number of countries, institutional or associational ones are a good deal more feasible.

All these remarks about means of linking up with outside resource centers and obtaining internationally published materials, however, miss one critical point, and it has to do with the constitution of evaluation

literature *within* the countries in question. Even if facilities for regular and full-fledged publication of evaluation-relevant texts or journals may not be available for some time to come, one sort of literature exists already and deserves to be much more fully collected and used to continuing professional education ends: the great variety of evaluations carried out over the years in the country in question and, if possible, in neighboring ones as well. Unfortunately, this material often is not much circulated within the country itself and, in some cases, it turns out five or ten years later that the only copies are to be found in the archives of international organizations or - supreme irony - in libraries of the North, now perhaps catalogued and digested for access via the Internet!

A first step of importance in remedying this situation, certainly feasible for an association, would lie therefore in simply collecting, circulating, comparing and storing the evaluation reports and findings developed locally. Let's remember that accumulation of precedents and records of clinical experience are one of the prime building blocks of any professional body of knowledge, and here is one of national origin that too often goes begging. As a second step then, associations might - perhaps with support from their Northern counterpart organizations or through collaboration with universities - begin publishing their own monograph series or journals.

Envoi: Optimism, patience and perseverance

These then are a series of suggestions for promoting - against all odds, one might say - the continuing professional education of evaluators and for helping them meet the challenge of involving and "educating" others in the evaluation endeavor. It is an effort very worth undertaking and supporting, even if, like a camel caravan, it takes a while to go the distance. Hopefully this book has furnished a few resources for the journey!

Notes

¹ See House (1993) for some valuable perspectives on "professionalizing" evaluation.

² Patton (1985) has sounded a similar note. It is less often remarked, but no less critical to remember, that existing cultures may have their own "traditional" modalities of information dissemination and of evaluation; and that evaluators will do well to inform themselves about these patterns and attitudes in the interests of couching participatory evaluation exercises in a language that will be easily understood and will "tap into" existing reflexes, while perhaps molding them to new ends. Judicious use of proverbs in African cultures can be one means of achieving this synergy of old and new, for there are many that deal with related topics and reveal local understandings of assessment and accountability. The topic is worth a chapter of its own!

³ And in this respect, at least, nonformal education and adult education evaluators are doubly favored. Since the field in which they work is already largely committed to principles and practices of "facilitation," "lifelong learning" and "self-direction," they should be among the best at adopting and perfecting this new mode of professionalism!

⁴ In academic parlance, "sabbatical" is used for a year off intervening in theory once every seven years, but in reality the scheduling is highly variable. Whatever the interval, the principle of "down time" or - as the French say - "resourcing" remains valid.

References

- Bhola, H. 1990. *Evaluation "Literacy for Development" Projects, Programs, and Campaigns: Evaluation Planning, Design and Implementation, and Utilization of Evaluation Results*. Germany: UNESCO Institute for Education.
- House, E.R. 1993. *Professional Evaluation: Social Impact and Political Consequences*. Newbury Park, CA: Sage Publications.
- Patton, M.Q. 1985. *Culture and Evaluation*. (New Directions for Program Evaluation 25). Newbury Park, CA: Sage.
- Schon, D.A. 1983. *The Reflective Practitioner: How Professionals Think in Action*. New York: Basic Books.
- Senge, P. 1990. *The Fifth Discipline: The Art and Practice of the Learning Organization*. New York: Doubleday.

Index

—A—

absorptive capacity (of meaning patterns) 169
accountability, 67, 295, 221, 303, 306, 307
and TQM, 210-213
administrative evaluation, 207-224
administrative model, 217-218
administrative output, 208-213
defining administrative objectives, 215-217
participatory aspects, 215-217, 219, 220, 222
professional bureaucracy, 219
team organization, 218-219
visioning, 215-217
administrative staff, 209
adult education, 21, 29, 34, 58, 67, 312
attribute variability, 136-138

—B—

balance sheet (for participatory program evaluation), 70, 75-80
in self-directed learner assessment, 226, 234, 239

—C—

causal association, 99, 251-257, 262
choreography of participatory evaluation, 40-46
CIPP/CIPOO model, 57, 242
constructivism, 86
constructivist tradition and evaluation, 87, 92, 94, 100
contextual understanding, 38-40, 156, 163, 164, 172, 175, 183-184, 194

continuing professional education, 3-4, 28, 303-315
correlation, 251, 253, 254, 255, 262
and causation in evaluation, 262
cost-benefit analysis, 75, 260, 286-294
cost-benefit evaluation, 286-294
criterion-referenced testing, 229, 239, 240
critical theory, 86, 94, 100, 102
cross-functional teams, 219
cross-tabulation tables, 253

—D—

data and their quality, 85-107
data patterns, 100-105
qualitative data and meaning, 159-163
qualitative data collection, 183-187
quantitative data definition, 107-109
decentralization, 23
decentralized administration, 207, 220
decentralized evaluation designs, 23-25
decentralization movement, 222
democratization, 23, 210
dimensions of participation, 31, 42-46
discursive validity, 92
divergent and convergent phases, 185-186
domains of evaluation systems model, 59-61, 207-208, 224, 243-245
double-loop thinking, 40

—E—

- ecological fallacy, 136
- empirical consequence (evaluation axis), 79
- empowerment, 6, 28, 29, 30, 38, 208, 215, 242, 263, 294
- ethnography, 163, 165, 169, 185-187, 221
- evaluability assessment, 245
- evaluation
 - formative, 3, 16, 20, 22, 122-124
 - macro and micro designs, 19-22
 - participatory, 28-55
 - summative, 3, 16, 123
- evaluation design, 13-27
 - centralized and decentralized designs, 23
 - concentric principle, 46, 237
 - evaluating the evaluation, 49
 - external communication, 45
 - internal and external evaluation, 48
 - lateral communication, 44-46
 - learning process approach, 14-18, 23
 - product and process dimensions, 22-23
 - vertical communication, 42-44
- evaluation research, 20
- experimental research, 249, 253, 255
- external customers, 217, 218, 219

—F—

- feasibility studies, 259-260
 - in nonformal education, 260

—G—

- general indicators, 111
- generalizability of data, 135, 136, 138, 145
 - in qualitative studies, 152
- goal displacement, 248
- goal-free evaluation, 68
- good indicators, 110-112
- grounded theory, 171, 192

—H—

- heuristic, 171
- hypothesis testing, 99

—I—

- impact evaluation, 242-265
 - counter-factual question, 245
 - defining program effects, 245
 - mixed-method approaches, 256-257
 - narrowing the focus, 246-251
 - program strategy, 244-245
 - qualitative or naturalistic approaches, 251-254
 - quantitative or rationalistic strategies, 253-256
 - selection criteria for impact focus, 245
 - tracer studies, 256-257, 257
 - unanticipated affects, 247-249
- implementation process, 68, 122, 162-163, 244-245, 251
- indicators, 107-129
 - direct, 113, 114
 - indirect, 113, 114
 - multiple, 119
 - outcome, 121
 - process, 120, 121, 122

proxy, 113, 114
 qualitative, 116
 quantitative, 113-115
 refining of, 112
 standardization of categories,
 117
 triangulation of, 112, 120
 indigenous models, 69-70
 information generation matrix,
 123-126
 instrument development, 31, 87,
 94, 112, 125
 and reliability of measurement,
 113
 internal customers, 217, 218, 219
 intersectoral collaboration, 258,
 262-265, 308
 interviewing, 164, 174-175, 185-
 190, 193, 194, 201

—K—

key informants, 135, 165, 177,
 185, 188, 191

—L—

learner assessment, 225-241
 learning organizations, 16, 309
 literacy, 225-237, 247, 257, 258,
 260
 local knowledge, 35, 36, 69
 logical consequence (evaluation
 axis), 79

—M—

management by the objectives
 (MBO), 44, 210
 management information system
 (MIS), 22, 123-124, 259

evaluation "mosaics" and de-
 centralized MIS, 126-127
 margin of error, 90-96, 104, 106,
 134, 136, 139-140, 144, 147,
 152, 236
 and reliability concerns, 88
 and representativity concerns,
 89-90
 and validity concerns, 90-92
 meaning patterns, 163, 166-173,
 191, 200, 204
 member checks, 187, 188, 200-
 202
 mental schemata, 166
 monitoring, 20
 motivational potential, 169

—N—

needs assessment, 76, 243
 non-governmental organizations
 (NGOs), 24

—O—

operational indicators, 111
 orders of magnitude, 97-99

—P—

participatory administrative eval-
 uation, 213-215, 222
 participatory evaluation design, 33-
 52
 and technical expertise, 36
 choreography, 40-46
 concentric principle, 46
 participatory evaluation, 28-55
 participatory evaluation methods,
 33-55, 225-241
 participatory learner assessment,
 225-241

patterns of data, 100-105
 pattern recognition, 101
 population (in sampling), 132, 134
 census of population units, 155
 defined, 135-136
 variability within, 137
 positivism, 85, 87
 post-positivism, 85, 87
 privatization, 23
 productive potential (of meaning patterns), 169
 program model, 63, 70, 73, 75, 110, 123-125, 127, 155, 209, 215, 248-250
 program theory, 38, 56-82, 108, 151, 170, 195, 209, 242, 245-251
 program-as-implemented, 76, 79
 program-as-planned, 76, 79
 programs as meaning patterns, 170-171
 project development, 13
 project/program implementation, 37, 75, 76, 109, 123-124, 143, 160-162, 220, 244, 251

—Q—

qualitative and quantitative indicators, 116-120
 qualitative evaluation
 analyzing and understanding data, 198-201
 coding and grouping data, 198
 collecting and managing data, 183-189
 controlling bias, 184
 data analysis, 189-191
 design considerations, 173-175
 ethical dilemmas, 188
 memos, 189

obtaining entry, 184
 synthesizing data, 200
 tape recording of, 187
 thick description, 163-166
 transcribing data, 186
 types of inquiry, 173-175
 use of teams, 187
 validation, 200
 qualitative evaluation methods, 159-206
 case studies, 173-174
 cultural and social "algorithms", 167-168
 discovering themes and categories, 195-197
 documentary content analysis, 175
 in-depth interviewing, 172
 meaning patterns, 161, 164-171, 189, 198, 202
 objectivity, 166-169
 participant observation, 172
 thick description, 160, 162, 163
 use of artifacts, 175
 qualitative sampling strategies, 134, 148-153
 quantitative data, 89-99
 quasi-experimental research, 253 ff.

—R—

reconstituting program theory, 70-75
 reflective practitioner, 14, 308
 reification, 102
 reliability of data, 87, 90-91, 93, 94, 95, 99, 101, 109, 111, 113, 115-116, 118, 127
 representativity of data, 92-94, 95, 104-105, 113

Index

323

research and development (R&D),
21, 52
Russian doll effect, 63

—S—

sample size, 138-148
sampling
 cluster, 146
 demographic, 131
 elite, 150, 177
 group, 151
 intentional or purposive, 149
 kinds of, 131
 methodological, 132
 mixed methods, 153-154
 multi-stage, 146
 nonprobability samples, 133,
 152
 probabilities proportional to
 size (PPS), 143
 qualitative, 148-153
 quantitative, 133-148
 random, 140-142
 sample size, 138-148
 snowball, 149
 stratified random, 142-146
 temporal, 131, 154
 thematic, 131, 155
 theoretical, 131, 156
sampling interval, 141
self directed learner assessment,
225-241
 cross-validation of, 236-237
 level of accuracy, 236
 reliability of, 235
 self evaluation instruments,
 230-234
sensitivity analysis, 99
significant figures, 96-97

site-based decision making
(SBDM), 210
stakeholders/stakeholder groups,
13-14, 18, 19, 22, 40, 41, 46,
52-53, 69, 80, 87, 88, 100-101,
155, 174, 185, 198, 200, 202,
203, 216, 248
statistically representative, 132,
135
systems model, 57-59
systems modeling of programs, 57-
67
systems thinking, 7, 38, 57, 119,
207

—T—

total quality management, 49, 210-
213
traditional knowledge, 35
triangulation, 89, 102, 119, 157

—U—

units of analysis, 135

—V—

validation, 202
 by member checks, 202

—W—

workable truth, 100

The capacity to evaluate education programs and draw lessons from their successes and failures is an essential component of educational development. *Sharpening our Tools* is an important contribution to this vital area. Addressed to evaluators of adult and nonformal education programs who already possess the basic skills of that profession, it takes a range of tools for evaluation and hones them to a fine edge. It discusses the applications of the tools, the various alternative designs and what to expect of them. It also encourages the reader to think more broadly and deeply about the evaluation process itself, its aims and its theoretical underpinnings.

BEST COPY AVAILABLE

352